

# ANÁLISIS ECOLÓGICO DE LA OCURRENCIA DE CASOS DE HIDATIDOSIS EN LA PROVINCIA DE RÍO NEGRO

*Por Marcos Albano Arezo*

*Presentado ante la Facultad de Matemática, Astronomía y Física  
y el Instituto de Altos Estudios Espaciales Mario Gulich  
como parte de los requerimientos para la obtención del grado de*

MAGÍSTER EN APLICACIONES ESPACIALES DE ALERTA  
Y RESPUESTA TEMPRANA A EMERGENCIAS

UNIVERSIDAD NACIONAL DE CÓRDOBA

Mayo, 2016

©IG - CONAE/UNC 2015

© FAMAFA - UNC 2015

**Director:** Dr. Edmundo Larrieu<sup>1,2</sup>  
**Co Director:** Lic. Mario Lanfri<sup>3</sup>

<sup>1</sup> Ministerio de Salud, Provincia de Río Negro, Argentina

<sup>2</sup> Facultad de Ciencias Veterinarias, Universidad Nacional de Río Negro, Argentina

<sup>3</sup> Instituto de Altos Estudios Espaciales Mario Gulich, Comisión Nacional de Actividades Espaciales (CONAE). Argentina.



Análisis Ecológico de la ocurrencia de casos de hidatidosis en la Provincia de Río Negro por Marcos Albano Arezo se distribuye bajo una [Licencia Creative Commons Atribución-NoComercial-CompartirIgual 2.5 Argentina](https://creativecommons.org/licenses/by-nc-sa/2.5/arg/).

---

## Resumen

La Hidatidosis, es una zoonosis parasitaria, endémica en la provincia de Río Negro. Las áreas geográficas de mayor riesgo están ubicadas al oeste y al centro de la provincia, en donde las condiciones ecológicas favorecen la sobrevivencia de huevos de *Echinococcus granulosus* y las condiciones sociales, culturales y económicas generan un ambiente epidemiológico que favorecen el ciclo de transmisión.

En este estudio se evalúa la relación entre la distribución geográfica de la tasa de hidatidosis en niños y factores de riesgo ambientales y climáticos como temperatura, vegetación y regiones naturales. Para tal fin se emplearon modelos espaciales de Poisson de efectos fijos y mixtos.

La temperatura media anual de la superficie terrestre y el Índice de Vegetación de Diferencia Normalizada, influyeron en forma negativa, y la región ecológica Patagonia Extra Andina Occidental en forma positiva, sobre la razón de morbilidad estandarizada de hidatidosis en chicos de 0 a 14 años. Cuando se evalúa la temperatura por rangos, la correlación se obtuvo con el intervalo de 17.19 a 20.90 °C.

Las covariables mencionadas explicaron en parte la distribución de casos en el área de estudio mediante el modelo de efectos fijos, presentando sobredispersión. Modelos jerárquicos Bayesianos espaciales de efectos mixtos modelaron la sobredispersión mediante la incorporación de efectos aleatorios estructurados y no estructurados en el espacio, favoreciendo la bondad de ajuste y el potencial de predicción del modelo.

La descomposición del riesgo relativo indicó que los radios censales ubicados al centro-sur y sur-oeste provincial fueron explicados por efectos fijos, mientras el riesgo relativo del radio censal n° 4, ubicado en el centro-norte, es debido a efectos aleatorios no estructurados en el espacio.

Los modelos desarrollados podrían ser utilizados para predecir la distribución de casos en el área programa Hidatidosis.

**Palabras claves:** equinococosis quística, mapeo de enfermedades, Análisis Ecológico, Modelos jerárquicos Bayesianos Espaciales, sobredispersión, efectos aleatorios espacialmente correlacionados y no correlacionados en el espacio.

---

## Abstract

Echinococcosis, a parasitic zoonosis is endemic in the province of Río Negro. The geographic areas of highest risk are located to the west and center of the province, where ecological conditions favor the survival of *Echinococcus granulosus* eggs and social, cultural and economic conditions generate an epidemiological environment that promote the sustenance of the transmission cycle.

In this study the relationship between the geographical distribution of the cases of hydatid disease and environmental risk factors such as temperature and climate, vegetation, rainfall, humidity and natural regions is evaluated. To this purpose Poisson spatial models of fixed and mixed effects were used.

The average Land Surface Temperature and Normalized Difference Vegetation Index, negatively influenced, and ecological region Extra Andean Patagonia Occidental positively on standardized morbidity rate of hydatid disease in children of 0-14 years. When assessing ranges, Land Surface Temperature from 17.19 to 20.90 °C are positively correlated with Standardized Morbidity Rate.

The covariates mentioned partly explained the distribution of cases in the study area using the fixed effects model, presenting overdispersion. Spatial hierarchical Bayesian models of mixed effects overdispersion modeled by incorporating structured and unstructured random effects in space, favoring the potential goodness of fit and model prediction.

The decomposition of relative risk indicated that the census radios located at the south-central and south-west provincial were explained by fixed effects, while the relative risk of the number 4 census radius in the center-north is due to random effects unstructured in space.

The developed models could be used to predict the distribution of cases of hydatid disease in the area program.

**Keywords:** cystic echinococcosis, Disease mapping, Ecological Analysis, Hierarchical Bayesian Spatial Models, overdispersion, uncorrelated and spatially correlated random effects.

---

## *Agradecimientos*

- *A mi esposa Natalia e hijo Máximo, por haber cedido de su tiempo y acompañarme siempre.*
- *A mi profe Eddy Larrieu, quien me introdujo en la materia.*
- *A mi tutor de maestría Mario Lanfri, quien compartio todo su conocimiento.*
- *A mis profesores MAEARTE, a la gente del Instituto de Altos Estudios Espaciales Mario Gulich, a la Universidad Nacional de Córdoba y a la Comisión Nacional de Actividades Espaciales.*
- *A mis compañeros de maestría, como se extrañan!*
- *A mis compañeros de trabajo de Salud Ambiental del Ministerio de Salud de Río Negro, quienes trabajan desde hace años en la vigilancia, prevención y control de la Hidatidosis!*
- *Al Tribunal evaluador: Dr. Scavuzzo, Dra. Konigheim y Dra. Provensal.*

# CONTENIDO

## INTRODUCCIÓN

<b>Capítulo 1:</b> -----	<b>1</b>
<b>1. Presentación del problema, hipótesis, objetivos y justificación.</b> -----	<b>1</b>
<b>1.1. Presentación del problema y motivación.</b> -----	<b>1</b>
<b>1.2. Hipótesis</b> -----	<b>4</b>
<b>1.3. Objetivos</b> -----	<b>4</b>
1.3.1. Objetivos generales-----	4
1.3.2. Objetivos específicos-----	5
<b>1.4. Justificación.</b> -----	<b>5</b>
<b>1.5. Presentación.</b> -----	<b>5</b>

## MARCO TEÓRICO

<b>Capítulo 2:</b> -----	<b>7</b>
<b>2. Hidatidosis</b> -----	<b>7</b>
<b>2.1. Biología de <i>Echinococcus granulosus</i></b> -----	<b>7</b>
<b>2.2. Epidemiología</b> -----	<b>9</b>
<b>2.3. Control y vigilancia de la equinococcosis</b> -----	<b>11</b>
<b>Capítulo 3:</b> -----	<b>13</b>
<b>3. Sistemas de información Geográfica, Sensado Remoto y Epidemiología Espacial.</b> -----	<b>13</b>
<b>3.1. Sistemas de Información Geográfica</b> -----	<b>13</b>
<b>3.2. Sensado Remoto</b> -----	<b>14</b>
<b>3.3. Epidemiología Espacial y Análisis Ecológico</b> -----	<b>17</b>
<b>Capítulo 4:</b> -----	<b>19</b>
<b>4. Mapeo de enfermedades y Modelos jerárquicos Bayesianos</b> -----	<b>19</b>

<b>4.1. Mapeo de enfermedades Bayesiano</b> -----	<b>19</b>
4.1.1. Representación bruta de la distribución de la enfermedad -----	21
4.1.2. Razón de mortalidad/morbilidad estandarizada-----	21
<b>4.2. Modelos jerárquicos Bayesianos</b> -----	<b>22</b>
4.2.1. Verosimilitud y Distribución a posteriori-----	23
4.2.2. Modelos Jerárquicos -----	23
4.2.3. Inferencia a Posteriori-----	24
4.2.4. Métodos Markov Chain Monte Carlo-----	25
<b>4.3. Modelo de Poisson para conteo de datos.</b> -----	<b>27</b>
<b>4.4. Sobredispersión</b> -----	<b>31</b>
<b>4.5. Modelo Log-Normal</b> -----	<b>32</b>
<b>4.6. Modelos de heterogeneidad correlacionada</b> -----	<b>32</b>
<b>4.7. Modelo Poisson para conteos de datos con ambos efectos aleatorios (UH y CH).</b> -----	<b>33</b>
<b>4.8. ANÁLISIS ECOLÓGICO</b> -----	<b>35</b>
<b>4.9. Residuos y bondad de ajuste</b> -----	<b>38</b>
4.9.1. Medidas de bondad de ajuste para modelos-----	38
4.9.2. Residuos generales-----	43

## **METODOLOGÍA Y RESULTADOS**

<b>Capítulo 5:</b> -----	<b>47</b>
--------------------------	-----------

### **5. Metodología y análisis e interpretación de los resultados.-47**

<b>5.1. Región de estudio</b> -----	<b>47</b>
<b>5.2. Definición y construcción de variables</b> -----	<b>47</b>
5.2.1. Covariables-----	48
5.2.2. Selección de covariables -----	51
<b>5.3. Análisis exploratorio de los datos</b> -----	<b>52</b>
<b>5.4. Modelos Jerárquicos Bayesianos</b> -----	<b>53</b>
5.4.1. Modelo espacial de EF (Regresión de Poisson)-----	53
5.4.2. Residuos y autocorrelación-----	56
5.4.3. Sobredispersión-----	58
5.4.4. Modelos espaciales de efectos Mixtos (modelo log-normal)-----	58

## **CONCLUSIONES**

<b>Capítulo 6:</b> -----	<b>68</b>
--------------------------	-----------

### **6. CONCLUSIONES ----- 68**

## Referencias. ----- 70

## Apéndice A: Figuras.----- 90

Figura 1. Esquema del ciclo de transmisión de <i>Echinococcus granulosus</i> .-----	90
Figura 2. Muestra la relación entre la resolución espacial, temporal y espectral de sensores satelitales usados o de potencial uso en epidemiología. -----	91
Figura 3. Región de estudio, provincia de Río Negro. -----	91
Figura 4. Mapa de casos de hidatidosis en chicos de 0 a 14 años, provincia de Río Negro. 92	
Figura 5. Mapa de rangos de temperatura media anual (LST) en grados Celsius. -----	92
Figura 6. Mapa de temperatura media anual (LST) para el rango de 17.19 a 19.90 °C y otros rangos. -----	93
Figura 7. Mapa de NDVI medio anual.-----	93
Figura 8. Mapa de la región ecológica PEAO y otras regiones. -----	94
Figura 9. Densidad de distribución de la SMR de hidatidosis en chico de 0 a 14 años. -----	94
Figura 10. Densidad de distribución del logaritmo de la SMR de hidatidosis en chicos de 0 a 14 años. -----	95
Figura 11. Histograma de SMR con curva normal superpuesta.-----	95
Figura 12. Histograma de log(SMR) con curva normal superpuesta. -----	96
Figura 13. Gráfico del porcentaje del logaritmo de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus temperatura media anual (eje x). -----	96
Figura 14. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus rangos de temperatura media anual (eje x). -----	97
Figura 15. Gráfico del porcentaje del logaritmo de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus rangos de temperatura media anual (eje x). -----	97
Figura 16. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus rango de temperatura media anual (eje x). -----	98
Figura 17. Gráfico del logaritmo del porcentaje de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus rango de temperatura media anual (eje x). -----	98
Figura 18. Gráfico del porcentaje del logaritmo de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus NDVI medio anual (eje x). -----	99
Figura 19. Gráfico del logaritmo del porcentaje de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus regiones ecológicas (eje x). -----	99
Figura 20. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus temperatura media anual (eje x), distinguida por rangos de temperatura. -----	100

<b>Figura 21. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus temperatura media anual (eje x) y distinción en regiones ecológicas PEAO y otras.</b> -----	<b>100</b>
<b>Figura 22. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus NDVI medio anual (eje x) y distinción en regiones ecológicas PEAO y otras.</b> -----	<b>101</b>
<b>Figura 23. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus temperatura media anual (eje x) y distinción por presencia y ausencia de casos de hidatidosis en chicos de 0 a 14 años.</b> -----	<b>101</b>
<b>Figura 24. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus NDVI medio anual (eje x) y distinción por presencia y ausencia de casos de hidatidosis en chicos de 0 a 14 años.</b> -----	<b>102</b>
<b>Figura 25. Se muestra el mapa de casos observados (superior), casos predichos por el modelo (intermedio) y residuos (inferior), obtenidos por la diferencia de los anteriores, para la región de estudio y por radio censal, para el modelo MOD_1.</b> -----	<b>103</b>
<b>Figura 26. Moran's I scatterplot</b> -----	<b>104</b>
<b>Figura 27. Radios censales Outlier: 4, 15, 37 y 43.</b> -----	<b>104</b>
<b>Figura 28. Indicadores locales de asociación espacial para los residuos de MOD_1.</b> -----	<b>105</b>
<b>Figura 29. Mapas de casos observados (superior), casos predichos por el modelo (intermedio) y residuos (inferior) para la región de estudio y por radio censal, para el modelo MOD_3.</b> -----	<b>106</b>
<b>Figura 30. Boxplot mostrando el intervalo de credibilidad del 95% para la estimación del riesgo relativo de los efectos de la temperatura media anual de la superficie terrestre (Grados Celsius).</b> -----	<b>107</b>
<b>Figura 31. Boxplot mostrando el intervalo de credibilidad del 95% para la estimación del riesgo relativo de los efectos del NDVI.</b> -----	<b>107</b>
<b>Figura 32. Boxplot mostrando el intervalo de credibilidad del 95% para la estimación del riesgo relativo de los efectos de la región ecológica PEAO.</b> -----	<b>108</b>
<b>Figura 33. SMR de hidatidosis en niños de 0 a 14 años (superior), el RR fijado por el modelo de EM (intermedio) y los residuos del RR, representados por radio censal para la región de estudio (inferior).</b> -----	<b>109</b>
<b>Figura 34. Componentes de la variación de la morbilidad de Hidatidosis en chicos de 0 a 14 años.</b> -----	<b>110</b>
<b>Figura 35. Moran's I scatterplot del modelo de EM con efectos aleatorios incorporados (MOD_3).</b> -----	<b>111</b>
<b>Figura 36. Mapa de residuos del modelo MOD_3, mostrando los valores atípicos.</b> -----	<b>111</b>
<b>Figura 37. Indicadores locales de asociación espacial para los residuos de MOD_3.</b> -----	<b>112</b>
<b>Figura 38. Casos observados vs. Casos predichos por el modelo (MOD_3).</b> -----	<b>113</b>
<b>Figura 39. Residuos estandarizados vs. Casos predichos.</b> -----	<b>113</b>

## Apéndice B: Tablas ----- 114

Tabla 1:-----	114
<i>Código OpenBUGS para el modelo de covarianza bayesiano jerárquico de efectos fijos con 3 covariables</i> -----	114
Tabla 2:-----	115
Código Open BUGS para el modelo de covarianza bayesiano jerárquico de efectos mixtos con 3 covariables, con efectos aleatorios no estructurados.-----	115
Tabla 3:-----	116
<i>Código Open BUGS para el modelo de covarianza bayesiano jerárquico de efectos mixtos con 3 covariables, con efectos aleatorios no estructurados y estructurados.</i> -----	116
Tabla 4:-----	117
Pasos para la selección de covariables con el software R.-----	117
TABLA 5: -----	118
<i>Regresión de Poisson con el software R.</i> -----	118
TABLA 6: -----	119
<i>Resultados de la regresión de Poisson, procedimiento en el software R: Model1.</i> -----	119
TABLA 7: -----	120
<i>Resultados de la regresión de Poisson, procedimiento en el software R: Model2.</i> -----	120
TABLA 8: -----	121
<i>Procedimiento en el software R: Creación del modelo</i> -----	121
TABLA 9: -----	122
<i>Procedimiento en el software R: datos y parámetros</i> -----	122
TABLA 10: -----	123
<i>Comparación de los coeficientes de regresión de los modelos de EF computados usando estadística frecuentista (en R) y Bayesiana (en OpenBUGS).</i> -----	123
TABLA 11: -----	123
<i>Resumen de estadísticos del modelo de EF Bayesiano.</i> -----	123
TABLA 12: -----	123
<i>Procedimiento en el software R para el cálculo del estadístico Moran's I para los residuos.</i> -----	123
TABLA 13: -----	124
<i>Procedimiento en R para el estadístico de Moran's I para cuantificar la autocorrelación espacial en áreas.</i> -----	124
Tabla 14:-----	125

<i>Procedimiento en el software R: Creación del modelo (MOD_3: modelo de EM con efectos aleatorios estructurados y no estructurados).</i> -----	125
Tabla 15:-----	127
<i>Procedimiento en el software R: datos y parámetros.</i> -----	127
Tabla 16:-----	128
<i>Procedimiento en el software R: estandarización y centrado de variables.</i> -----	128
Tabla 17:-----	128
<i>Resultados de los modelos de EF (MOD_1), modelo de EM con efectos aleatorios no estructurados (MOD_2) y modelo de EM con efectos aleatorios estructurados y no estructurados (MOD_3).</i> -----	128
Tabla 18:-----	129
<i>Bondad de ajuste para los diferentes modelos.</i> -----	129
Tabla 19:-----	130
<i>Coefficientes de regresión para las covariables estandarizadas y centradas, y varianza de los componentes aleatorios no estructurados y estructurados (sd.u y sd.s respectivamente) de los diferentes modelos.</i> -----	130
Tabla 20:-----	131
<i>Modelo de EF y EM, usando como covariable la temperatura media anual de la superficie terrestre en forma binaria (valor = 1, para los radios censales que están en el rango de temperatura de 17.19 a 20.90 °C y valor = 0, para el resto).</i> -----	131
Tabla 21:-----	132
<i>Análisis de sensibilidad usando diferentes valores para la distribución Gamma del parámetro varianza inversa de los efectos aleatorios.</i> -----	132
Tabla 22:-----	132
<i>Resultados del test de Moran's I para el modelo de EM con efectos aleatorios incorporados (MOD_3).</i> -----	132

## **Apéndice C: Uso combinado de R y OpenBUGS: Breve descripción ----- 133**

## **Apéndice D: Procedimiento en el software R: evaluación de la salida del muestreo de Gibbs con CODA. ----- 134**

# LISTA DE ABREVIATURAS

- AIC:** criterio de información de Akaike
- BIC:** criterio de información bayesiano
- CAR:** modelos condicionales autorregresivos
- CH:** heterogeneidad correlacionada
- DIC:** criterio de información de desviación
- EF:** efectos fijos
- EM:** efectos mixtos
- EQ:** equinocosis quística
- GLM:** modelos lineales generalizados
- LISA:** indicadores locales de asociación espacial
- LST:** temperatura superficial de la tierra
- MA:** media móvil
- MAPE:** error predictivo absoluto medio
- MBA:** medidas de bondad del ajuste
- MC:** error estándar de Monte - Carlo
- MCMC:** métodos de cadenas de Markov Monte Carlo
- MSE:** error cuadrático medio
- MSPE:** error predictivo cuadrático medio
- MRT:** *MODIS Reprojection Tool*
- NAN:** *No data*
- NDVI:** índice de vegetación de diferencia normalizada
- NIR:** longitud de ondas infrarrojas cercanas
- PB:** *bootstrap* paramétrico
- pD:** número efectivo de parámetros
- PEAO:** Patagonia Extra Andina Occidental
- R:** longitudes de ondas rojas
- RR:** riesgo relativo

**SAR: autorregresivo simultáneo**

**SIG: Sistemas de Información Geográfica**

**SMR: razón de morbilidad estandarizada**

**SR: Sensado Remoto**

**UH: heterogeneidad no correlacionada**

## CAPÍTULO 1:

### 1. PRESENTACIÓN DEL PROBLEMA, HIPÓTESIS, OBJETIVOS Y JUSTIFICACIÓN.

#### 1.1. Presentación del problema y motivación.

La equinocosis quística (EQ) o hidatidosis es una zoonosis parasitaria producida por un cestodo familia *Taenidae*, género *Echinococcus*, especie *granulosus*. Requiere de dos hospederos mamíferos para completar su ciclo de vida. En América del Sur el hospedero definitivo de mayor importancia epidemiológica es el perro (donde se desarrolla la faz adulta o estrobilar) mientras que el hospedero intermediario de mayor importancia es el ovino (en donde se desarrolla la faz larvaria o metacestode), aunque cabras, cerdos y bovinos pueden estar involucrados en el ciclo de la enfermedad (Thompson y McManus, 2001, p. 9; Larrieu y Zanini, 2012) y donde el hombre juega el rol de huésped accidental al ingerir huevos fértiles adheridos al ano o pelos de perros parasitados, por la ingestión de verduras y aguas contaminadas con materia fecal canina (Larrieu et al., 2004) o suelos y ambientes contaminados.

América del Sur se encuentra entre las regiones más afectadas del mundo y en Argentina es una de las enfermedades zoonóticas de mayor prevalencia. Provoca elevados gastos para los sistemas de salud en razón de costos de internación y tratamiento de las personas, produciendo, además, pérdidas para la ganadería en función del valor de las vísceras decomisadas y pérdidas en la producción de lana, leche y carne (Budke, Deplazes y Torgerson, 2006).

La EQ es endémica en la Provincia de Río Negro. Enmarcado hoy en el Programa Nacional de control de Enfermedades Zoonóticas, el programa provincial de Hidatidosis, data de aproximadamente 30 años y se base fundamentalmente en la desparasitación de perros, educación para la salud, control de la faena (para garantizar el no acceso de perros a las vísceras) y legislación para regular las poblaciones caninas y definir la responsabilidad del Gobierno y ganaderos. Mientras que las acciones de vigilancia radican en la determinación de la tasa de infección en el perro mediante la detección de coproantígenos (Craig, 1997; Guarnera, Santillan, Botinelli y Franco, 2000; Pérez et al., 2006), identificación de ovinos parasitados mediante el análisis macroscópico en sala de faena (Larrieu et al., 2001; Cabrera et al., 2003), vigilancia de

la infección en el hombre mediante el registro de los casos humanos operados y mediante tamizajes serológicos (Coltorti y Varela-Díaz, 1978; Coltorti, 1986) y tamizaje ultrasonográfico para la determinación de la tasa de infección precoz en niños (Del Carpio et al. 2000; Frider y Larrieu, 2010).

El área endémica en donde se desarrolla el programa de control abarca 120013Km<sup>2</sup> incluyendo a los Departamentos de Bariloche, Pilcaniyeu, Ñorquinco, 25 de Mayo, 9 de Julio, Valcheta y El Cuy, en donde la producción ovina es la actividad ganadera más importante. La cría de ovejas y el número relacionado de perros de trabajo, agravado por la práctica de sacrificio de ovinos adultos para el consumo doméstico, permite a los perros alimentarse de vísceras infectadas, generando las condiciones ideales para la persistencia del ciclo de la enfermedad (Larrieu y Zanini, 2012).

Desde el punto de vista fitogeográfico, la Provincia de Río Negro comprende dos regiones diferenciadas: la estepa patagónica con su sucesión de mesetas escalonadas, escasas precipitaciones (menos de 300 mm anuales) y temperaturas extremas, y la región de los bosques andinos patagónicos, en la región cordillerana, al oeste de la provincia, caracterizadas por altas precipitaciones (más de 2500 mm anuales) y bajas temperaturas invernales.

Las áreas geográficas de mayor riesgo están ubicadas al oeste y al centro de la provincia (Larrieu, Lester, Jáuregui, Odriozzola y Aguero, 1986; Larrieu et al., 2000) incluyendo las localidades de Bolsón y Bariloche en la región cordillerana y las de Comallo, Pilcaniyeu, Ñorquinco e Ingeniero Jacobacci y sus áreas rurales en la región de la meseta patagónica, en donde las condiciones ecológicas favorecen la sobrevivencia de huevos de *Echinococcus granulosus* y las condiciones sociales, culturales y económicas generan un ambiente epidemiológico que favorecen el sustento del ciclo de transmisión (Larrieu, Del Carpio, Costa y Yadon, 2002).

Entre los factores que contribuyen a la dinámica de la transmisión de *Echinococcus granulosus* en establecimientos ganaderos, a parte de los factores socio-ecológicos e intrínsecos de los huéspedes, se encuentran los factores extrínsecos como la temperatura y humedad ambiental y los agentes que dispersan los huevos del parásito desde las heces al ambiente (Gemmell, Roberts, Beard y Lawson, 2001). Asimismo, la duración del efecto climático estacional en la supervivencia de los huevos es el que entre otras cosas determina la dinámica de transmisión y prevalencia geográfica (Gemmell et al., 2001).

En cuanto a la dispersión, la mayoría de los taenidos permanece dentro de los 180 metros del lugar de deposición, pero algunos pueden dispersarse rápidamente sobre un área de más de 30000 ha debido a la acción de factores climáticos como el viento y el agua, así como también por algunos animales como insectos y aves (Gemmell et al., 2001). Los huevos de *Echinococcus granulosus* son muy resistentes a las condiciones climáticas, pero, el clima cálido y seco, la heliofania prolongada y la radiación ultravioleta solar directa son los principales factores que inciden sobre su vitalidad (Guarnera, 2009). Incluso el cambio climático que provoca cambios de temperatura y

humedad del ambiente y/o suelo puede afectar la transmisión de *Echinococcus spp.* (Atkinson et al., 2013).

El ambiente juega un importante rol en la dinámica de las enfermedades y así se refleja al formar parte de uno de los componentes de la triada ecológica. La influencia de factores ambientales (físicos, químicos y biológicos) influye en las tasas de crecimiento de poblaciones de vectores de enfermedades infecciosas como insectos y roedores, como así también en la supervivencia de los huevos de *Echinococcus spp.* (Atkinson et al., 2013).

La descripción de las relaciones existentes entre la distribución geográfica de las enfermedades y los factores de riesgo ambientales y su análisis por medio de procedimientos estadísticos se denomina “Análisis Ecológico” (Kistemann, Dangendorf y Schweikart, 2002). Las características ambientales y climáticas más comúnmente usadas en estudios epidemiológicos de este tipo son, índice de vegetación de diferencia normalizada (NDVI), cobertura y uso del suelo, elevación, pendiente, aspecto, capas geológicas, presencias de cuerpos de agua, temperatura, precipitaciones y humedad (Rinaldi, Musella, Biggeri y Cringoli, 2006).

Los Sistemas de Información Geográfica (SIG) y el Sensado Remoto (SR) ofrecen medios eficaces para el mapeo de enfermedades, el análisis ecológico y la vigilancia epidemiológica y se han convertido en herramientas indispensables para el procesamiento, análisis y visualización de datos espaciales (Rinaldi et al., 2006).

Empleando estas herramientas y técnicas, una variedad de estudios han sido realizados en el campo de las enfermedades transmitidas por vectores (Reisen, 2010), principalmente en mosquitos (Arezo, 2012) y en Malaria (Machault et al., 2011). En el área de parasitología veterinaria véase (Cringoli, Rinaldi, Veneziano y Musella, 2005; Rinaldi et al., 2006) y para una revisión de los métodos y aplicaciones en epidemiología y ecología espacial de parásitos (Pullan, Sturrock, Soares Magalhães, Clements y Brooker, 2012), mientras que una descripción de modelos para helmintos transmitidos por el suelo puede véase en (Pullan et al., 2011).

El análisis de regresión se introducen como un medio para cuantificar el efecto de un conjunto de variables explicativas sobre la distribución espacial de una variable respuesta en particular (Pfeiffer, 2008). Cuando se cuenta con datos agrupados por área puede seguirse una aproximación estadística clásica o Bayesiana para este medio. En la segunda, modelos jerárquicos Bayesianos se han utilizado con el fin de cartografiar y estudiar la relación entre la variación geográfica de la enfermedad y covariables ambientales y socio-económicas. Estos modelos han sido aplicados en el campo de la salud humana, por ejemplo en cáncer (Clayton, Bernardinelli y Montomoli, 1993; Jarup, Best, Toledano, Wakefield y Elliott, 2002; Mohebbi, Wolfe, Jolley, Forbes, Mahmoodi y Burton, 2011; Puigpinós-Riera et al., 2011), mortalidad (Hoffmann et al., 2014; Sartorius, Sartorius, Chirwa y Fonn, 2011), diabetes mellitus insulino dependiente (Schober, Rami, Waldhor y Karimian-Teherani, 2001), en veterinaria para fasciola

hepática (Durr, Tait y Lawson, 2005) y lengua azul (Willgert, Schroedle y Schwermer, 2011), etc.

Una considerable literatura se ha desarrollado en torno a la aproximación bayesiana para modelar recuentos de la enfermedad a nivel de áreas pequeñas, véase a Mollié (1996), Best et al. (1999), Wakefield, Best, y Waller (2000), Lawson (2001), Banerjee, Carlin y Gelfand (2004, 2014), Waller and Gotway (2004), Waller (2005), Carlin y Louis (2008) y Lawson (2009, 2013). Central en este método es la inclusión de términos de efectos aleatorios para tener en cuenta características espaciales no observadas dentro de los datos. Lo atractivo de este enfoque es que en lugar de producir un coeficiente único (global) de autocorrelación espacial, como en el caso de los modelos condicionales autorregresivos (CAR), autorregresivo simultáneo (SAR) y media móvil (MA), términos de efectos aleatorios espacialmente correlacionados se estiman para cada unidad de superficie, lo que permite al analista identificar agregaciones de unidades espaciales donde la incidencia o prevalencia de la enfermedad no se explica por el modelo (Pfeiffer et al., 2008).

Generalmente se persiguen dos objetivos cuando se modelan datos de área. El primero es cuantificar la influencia de efectos fijos (EF) en la enfermedad para cada área o región de estudio y el segundo, identificar áreas donde hay conteos de casos elevados luego que la influencia de EF ha sido evaluada (Pfeiffer et al., 2008).

Debido a lo expuesto, es de interés conocer si existe influencia de factores climáticos y ambientales sobre la distribución geográfica de hidatidosis dentro del área programa Hidatidosis provincial. También es de interés investigar si existe exceso de enfermedad después de que la influencia de estos factores (fijos) ha sido evaluada, donde este exceso responde a factores de riesgo o variables no medidas en el estudio (efectos aleatorios). A una escala provincial, las variables no medidas podrían estar representadas en nuestro caso, por ejemplo, por factores de riesgo de índole social, cultural, productiva y económica.

## **1.2. Hipótesis**

VARIABLES CLIMÁTICAS Y AMBIENTALES COMO, temperatura, humedad, vegetación, precipitación, altura y regiones naturales, influyen en el ciclo biológico de *Echinococcus granulosus* determinando la distribución geográfica de la hidatidosis y en consecuencia influyen en el riesgo de enfermar para el hombre.

## **1.3. Objetivos**

### **1.3.1. OBJETIVOS GENERALES**

Determinar la relación entre la distribución geográfica de hidatidosis en niños y variables climáticas y ambientales.

### 1.3.2. OBJETIVOS ESPECÍFICOS

- Identificar áreas de prevalencia elevada (de casos de hidatidosis en chicos de 0 a 14 años) y explicar estas en términos de covariables ambientales y climáticas (temperatura, humedad, vegetación, precipitación, altura y regiones naturales)
- Conocer el exceso de enfermedad en cada área luego de que factores climáticos y ambientales han sido evaluados
- Representar en forma de mapas, el riesgo relativo en cada área, para EF y para efectos aleatorios.

### 1.4. Justificación.

El conocimiento de la distribución geográfica de Hidatidosis dentro del área programa es de ayuda para planificar y programar acciones de control y vigilancia epidemiológica. Conocer el rol que juegan el clima y los factores ambientales en dicha distribución permite mejorar el conocimiento de la mecánica de transmisión y de los factores de riesgo que posibilitan la ocurrencia de casos nuevos de esta enfermedad.

### 1.5. Presentación.

Habiéndose introducido previamente la problemática, hipótesis, objetivos y justificación de la tesis, en esta primera parte “Introducción”, se procede a una breve descripción de las otras dos partes restantes, a saber, Marco Teórico y Metodología, Análisis e Interpretación de los Resultados y Conclusiones.

En la segunda parte “Marco Teórico”, se integra el tema de investigación con las teorías, enfoques teóricos, estudios y antecedentes en general que se refieren al problema de investigación. Pueden distinguirse aquí tres capítulos, el capítulo 2, que trata de Hidatidosis, donde se describe la biología de *Echinococcus granulosus*, la epidemiología de la enfermedad y el control y la vigilancia, el capítulo 3, el que habla sobre SIG, SR y Epidemiología Espacial, y por último el capítulo 4, en el cual se describe el Mapeo de Enfermedades y los Modelos Jerárquicos Bayesianos, los que se desarrollan y emplean a lo largo de la tesis.

La última y tercera parte “Metodología, Análisis e Interpretación de los Resultados y Conclusiones”, consta de dos capítulos, el capítulo 5, donde se describe la metodología y se muestran e analizan los resultados, aquí se trata sobre la región de estudio, la definición y construcción de variables, el análisis exploratorio de los datos y los modelos jerárquicos Bayesianos. Por último, en el capítulo 6 se realizan las conclusiones, la que se enmarca en un contexto de inicio-cierre, que parte con las ideas propuestas o preliminares del estudio, luego describe los logros obtenidos, y finalmente formula otras ideas partiendo de las que se tuvieron al principio del estudio.

También puede verse al final, los apéndices A y B, de figuras y tablas respectivamente, y en apoyo a la metodología, un apéndice C, que describe brevemente el uso combinado del software R y OpenBUGS, y un apéndice D, donde se ve el procedimiento de evaluación de la salida del muestreo de Gibbs con el paquete CODA utilizando el software R.

## CAPÍTULO 2:

### 2. HIDATIDOSIS

#### 2.1. Biología de *Echinococcus granulosus*

##### AGENTE ETIOLÓGICO

La EQ o hidatidosis es una zoonosis parasitaria producida por un endoparásito perteneciente al Phylum Platyhelminthes, Clase Cestoda, familia *Taeniidae*, género *Echinococcus*, especie *granulosus*. Requiere de dos hospederos mamíferos para completar su ciclo de vida. Un hospedero definitivo, (carnívoro, especialmente el perro) donde se desarrolla la faz adulta o estrobilar y un hospedero intermediario en donde se desarrolla la faz larvaria o metacestode (Thompson y Mcmanus, 2001).

Existen diversidad de cepas de *Echinococcus granulosus*, algunas de las cuales se encuentran presente en Argentina: cepa ovina G1, cepa ovina tasmania G2, cepa vaca G5, cepa camello G6 y cepa porcina G7. (Eckert y Thompson, 1997; Rosenzvit et al., 1999; Kamenetzky et al., 2002; Cardona y Carmena, 2013; Carmena y Cardona, 2013).

##### FORMA ADULTA O ESTROBILAR

La forma adulta de *Echinococcus granulosus* corresponde a una tenia blanca que mide de 3 a 7 mm de longitud que se localiza en el intestino delgado del hospedador definitivo (principalmente el perro). Posee cuerpo segmentado y consiste en un número de unidades reproductivas o proglótides (generalmente 4), resultando la reproducción sexual y hermafrodita. Los proglótides maduros conteniendo un promedio de 587 huevos fértiles, son eliminados con la materia fecal del hospedador a razón de 0.071 proglótides por tenia por día. (Gemmel y Lawson, 1995; Thompson y Mcmanus, 2001).

El tiempo de vida del parásito se encuentra comprendido entre 10 meses y 4 años. (Thompson y Mcmanus, 2001).

##### HUEVOS DE ECHINOCOCCUS GRANULOSUS

Los huevos son de características ovoideas, de 30 a 40 micras de diámetro, conteniendo un embrión hexacanto (oncósfera o primer estado larval) envuelto en varias membranas y una gruesa pared keratinizada y de alta resistencia (embriósforo) (Thompson y Mcmanus, 2001).

Depositados en el ambiente, los huevos pueden llegar a desplazarse hasta 180 metros del lugar de defecación y pueden ser dispersados en áreas de hasta 30.000 ha por dípteros y escarabajos coprófagos que actúan como transportadores, contaminando grandes extensiones de campo, el agua de arroyos y pozos de bebida, verduras, etc., pudiendo también permanecer adheridos a los pelos y ano del perro. (Gemmel y Lawson, 1995; Gemmel et al., 2001)

Los huevos de *Echinococcus granulosus* son muy resistentes a las condiciones climáticas, pero, el clima cálido y seco, la heliofania prolongada y la radiación ultravioleta solar directa son los principales factores que inciden sobre su vitalidad (Guarnera, 2009). Pueden permanecer viables un año en un amplio rango de temperatura (4 a 15 grados centígrados). Son sensibles a la desecación pudiendo morir en 4 días a una humedad ambiente de 0% o en 5 días a una temperatura de 60° (Gemmel y Lawson, 1983, 1995; Gemmel et al., 2001).

#### **FORMA LARVAL O METACESTODE**

Cuando huevos de *Echinococcus granulosus* son ingeridos por hospederos susceptibles (especialmente el ovino) llegan al estómago en donde sufren la disgregación del embriósforo y la activación de la oncósfera. Esta exhibe intrincados movimientos rítmicos del cuerpo y los ganchos y penetran a través de las microvellosidades intestinales pasando a los sistemas linfáticos y venosos para ir a instalarse definitivamente en alguna víscera, preferentemente hígado o pulmón. (Thompson y Mcmanus, 2001)

La reorganización de la oncósfera y la formación de la membrana germinal y laminar ocurren rápidamente (en los primeros 14 días). El período prepatente es largo pudiendo tardar, según distintos autores, entre 2 y 6.5 años en producir protoescólices (Thompson y Mcmanus, 2001).

La presencia de protoescólices transforma al quiste hidatídico en fértil. La presencia de protoescólices vivos, además, clasifica al quiste hidatídico como viable (Cavagión, Alvarez y Larrieu, 2002; Yildiz y Gurcan, 2003).

#### **CICLO BIOLÓGICO**

Los carnívoros se infectan por la ingestión de protoescólices viables ubicados en las vísceras infectadas por la forma larval. Ovinos (*Ovis aries*) parasitados pueden ser faenados por el hombre con fines de alimentación o pueden morir en el campo por otras causas. Si sus vísceras son ingeridas por un perro o por otro carnívoro, los cientos de embriones hexacantos contenidos en un quiste hidatídico serán liberados en el intestino

del animal dando lugar a la formación de nuevas tenias (Gemmel y Lawson, 1995), ver Figura 1, apéndice A: figuras.

El período prepatente es corto, aproximadamente 7 semanas, momento en que comienza la liberación con la materia fecal del perro (*Canis lupus familiaris*) del primer proglótide maduro con su carga de huevos infectantes. (Gemmel y Lawson, 1995).

Los ungulados (ovinos, bovinos, porcinos, caprinos, guanacos, etc.) adquieren la infección por la ingestión de huevos del parásito liberados al ambiente (Thompson y Mcmanus, 2001).

El potencial biótico de *Echinococcus granulosus* (número potencial de quistes viables que pueden ser producidos por perro infectado por día) es bajo comparado con otros cestodes, por lo que la presión de infección en el ovino puede no ser suficiente para producir inmunidad adquirida por ingestión continua de huevos, especialmente en las dos semanas siguientes a la infestación. En áreas de alta prevalencia la presencia masiva de huevos podría generar inmunidad adquirida en los ovinos, la cual puede permanecer largo tiempo e impedir una nueva infestación o puede perderse cuando las áreas de pastoreo permanecen libres de huevos durante 6 a 12 meses, provocando un aumento de la prevalencia de la infestación con la edad. (Gemmel y Lawson, 1995; Cabrera et al., 1995)

Si bien el ciclo principal de *Echinococcus granulosus* comprende animales domésticos, pueden producirse ciclos selváticos entre el zorro (*Dusicyon culpaeus*, *Dusicyon griseus*, *Vulpes vulpes*) y la liebre (*Lepus europaeus*), (Schantz, Colli, Cruz-Reyes y Prezioso, 1976; Gemmel et al., 2001).

Las distintas cepas de *Echinococcus granulosus* pueden poseer períodos de prepatencia de diferente magnitud y diferente infectividad para el hombre. (Eckert y Thompson, 1997; Rosenzvit et al., 1999; Cardona y Carmena, 2013; Carmena y Cardona, 2013)

## **EQUINOCOCCOSIS QUÍSTICA EN EL HOMBRE**

El hombre es hospedero del metacestode y se infecta al ingerir huevos fértiles adheridos al ano o pelos de perros parasitados o por la ingestión de verduras o aguas contaminadas con materia fecal canina.

Los huevos eclosionan liberando el embrión hexacanto en el intestino delgado, pasan a la circulación venosa hasta alojarse en el hígado (principal localización), pulmón (segunda localización de importancia) u otra víscera o tejido, donde se formará la hidátide o quiste hidatídico (generalmente solo uno), (Ammann y Eckert, 1995).

### **2.2. Epidemiología**

#### **FACTORES DE RIESGO**

La EQ afecta principalmente a los habitantes de zonas rurales. Los principales factores de riesgo son la cría de lanares asociada a la tenencia de gran número de perros y al hábito de faenar ovinos adultos para consumo propio y alimentación del perro.

### **PÉRDIDAS ECONÓMICAS**

La EQ es una de las enfermedades zoonóticas que produce mayores pérdidas económicas en función del valor de las vísceras decomisadas, pérdidas en la producción de lana, leche y carne y para los sistemas de salud en razón de los costos de internación y tratamiento de las personas (Larrieu et al., 2000, Torgerson, 2003; Budke et al., 2006).

### **DISTRIBUCIÓN GEOGRÁFICA Y OCURRENCIA**

América del Sur se encuentra entre las regiones del mundo más afectada por la EQ y en Argentina es una de las enfermedades zoonóticas de mayor prevalencia.

La EQ es endémica en la Provincia de Río Negro, que cuenta con un programa de control que data de más de 30 años.

El área endémica en donde se desarrolla el programa de control abarca 120.013Km<sup>2</sup> incluyendo a los Departamentos de Bariloche, Pilcaniyeu, Ñorquinco, 25 de Mayo, 9 de Julio, Valcheta y El Cuy, en donde la producción ovina es la actividad ganadera más importante.

La cría de ovejas y el número relacionado de perros de trabajo, agravado por la práctica de sacrificio de ovinos adultos para el consumo doméstico, permite a los perros alimentarse de las vísceras infectadas, generando las condiciones ideales para la persistencia del ciclo de la enfermedad (Larrieu y Zanini, 2012).

Limitaciones geográficas, climáticas y de infraestructura afectan la posibilidad de efectuar visitas domiciliarias para la desparasitación canina en gran parte del año afectando las coberturas, cumpliéndose menos actividades que las programadas, de lo cual resulta el mantenimiento basal de la infección en perros y la ocurrencia de casos nuevos (Larrieu, comunicación personal).

Desde el punto de vista fitogeográfico, la Provincia de Río Negro comprende dos regiones diferenciadas: la estepa patagónica con su sucesión de mesetas escalonadas, escasas precipitaciones (menos de 300 mm anuales) y temperaturas extremas, y la región de los bosques andinos patagónicos, en la región cordillerana, al oeste de la provincia, caracterizadas por altas precipitaciones (más de 2500 mm anuales) y bajas temperaturas invernales.

Las áreas geográficas de mayor riesgo están ubicadas al oeste y al centro de la provincia (Larrieu et al., 1986, 2000) incluyendo las localidades de Bolsón y Bariloche en la región cordillerana y las de Comallo, Pilcaniyeu, Ñorquinco e Ingeniero Jacobacci y sus áreas rurales en la región de la meseta patagónica, en donde las condiciones ecológicas favorecen la sobrevida de huevos de *Echinococcus granulosus* y las condiciones

sociales, culturales y económicas generan un ambiente epidemiológico que favorece el sostenimiento del ciclo de transmisión (Larrieu et al., 2002; Craig y Larrieu, 2006; Larrieu y Zanini 2012).

#### **EFFECTOS DEL AMBIENTE Y EL CLIMA EN LA DINÁMICA DE TRANSMISIÓN.**

Entre los factores que contribuyen a la dinámica de transmisión de *Echinococcus granulosus* en establecimientos ganaderos, figuran los factores socio - ecológicos e intrínsecos de los huéspedes, factores extrínsecos como la temperatura y humedad ambiental y los agentes que dispersan los huevos del parásito desde las heces al ambiente (Gemmell et al., 2001).

La duración del efecto climático estacional en la supervivencia de los huevos, es el que entre otras cosas, determina la dinámica de transmisión y prevalencia geográfica (Gemmell et al., 2001).

El cambio climático puede provocar alteración en los valores de temperatura y humedad del ambiente y suelo, que impactan en la transmisión de *Echinococcus spp.*, a través de la supervivencia de los huevos en el medio externo (Atkinson et al., 2013).

El ambiente juega un importante rol en la dinámica de las enfermedades y así se refleja al formar parte de uno de los componentes de la triada ecológica. La influencia de factores ambientales (físicos, químicos y biológicos) influye en las tasas de crecimiento de poblaciones de vectores de enfermedades infecciosas como insectos y roedores, como así también en la supervivencia de los huevos de *Echinococcus spp.* (Atkinson et al., 2013).

### **2.3. Control y vigilancia de la equinococcosis**

#### **CONTROL**

Las actividades desarrolladas en los programas de control se basan en la desparasitación de perros con praziquantel (droga tenicida no ovicida); educación para la salud, control de la faena para garantizar el no acceso de perros a vísceras y legislación para la regulación de las poblaciones caninas y definición de responsabilidades de Gobierno y ganaderos.

#### **VIGILANCIA EPIDEMIOLÓGICA**

Los sistemas de vigilancia epidemiológica integrados a los programas de control han incluido tradicionalmente, Identificación de perros parasitados mediante dosificación con el tenífugo bromhidrato de arecolina (Schantz, 1973), vigilancia de la infección en el hombre mediante el registro de los casos humanos operados y mediante tamizajes

serológicos (Coltorti y Varela-Díaz, 1978; Coltorti, 1986) y ultrasonográfico para la determinación de la tasa de infección precoz en niños (Del Carpio et al., 2000; Frider y Larrieu, 2010), identificación de ovinos parasitados mediante el análisis macroscópico en sala de faena (Larrieu et al., 2001; Cabrera et al., 2003) y determinación de la tasa de infección en el perro mediante la detección de coproantígenos (Craig y Larrieu, 2006; Guarnera et al., 2000; Perez et al., 2006; Frider y Larrieu, 2010).

## CAPÍTULO 3:

### 3. SISTEMAS DE INFORMACIÓN GEOGRÁFICA, SENSADO REMOTO Y EPIDEMIOLOGÍA ESPACIAL.

#### 3.1. Sistemas de Información Geográfica

Los SIG han sido definidos de diferentes maneras, pero, en general, un SIG es una plataforma que consta de hardware, software, datos y recurso humano y comprende un fundamental y universalmente aplicable set de herramientas de valor agregado para la captura, transformación, gestión, análisis y presentación de la información geográficamente referenciada (geo-referenciada).

Los datos digitales dentro del SIG se pueden presentar en forma de mapa usando los llamados capas de datos que representan la información recopilada.

Dos enfoques pueden ser utilizados, el modelo de datos vectorial y el modelo de datos raster. En un modelo de datos vectoriales, todas las líneas están representadas por cadenas de vectores, y las áreas por polígonos. Los atributos se codifican en tablas separadas utilizando caracteres alfanuméricos como una etiqueta para una clase específica o categoría de propiedades. El modelo de datos raster utiliza una red de polígonos adyacentes (denominados "células") para proporcionar una cubierta virtual de una parte determinada de un territorio. Las células son a menudo llamados píxeles, y valores de atributo de los objetos que las células representan se asigna a los píxeles correspondientes (Daniel, Kolár y Zeman, 2004). Los datos ráster se utilizan normalmente para representar fenómenos continuos, por ejemplo, mapas de cobertura de la tierra, modelos digitales de elevación y mapas de fenómenos climáticos.

En el campo de la epidemiología geográfica, SIG se utilizan para la elaboración de mapas de enfermedad y para el análisis ecológico.

En el marco del análisis ecológico, SIG puede ayudar de manera significativa con la evaluación de la distribución de los factores ambientales relevantes en salud mediante la interpolación y el modelado (Kistemann, Dangendorf y Schweikart, 2002).

SIG puede ayudar contestando algunas preguntas importantes tales como "¿qué es?", "¿dónde está?" y ¿Cuándo ocurrió?, que es importante en la mayoría de los procesos ambientales y epidemiológicos (Rinaldi et al., 2006).

La capacidad de estos sistemas justifica su amplia aplicación que incluye las ciencias médicas veterinarias y humanas. Varios tipos de análisis utilizados rutinariamente en los SIG pueden ser muy útiles para la epidemiología espacial.

Estos incluyen, análisis de vecindad (todas las características que se ajusten a determinados criterios y son adyacentes a una característica particular, se encuentran y se enumeran), generación de búfer (generación de zonas de amortiguación alrededor o a lo largo de ciertas características), análisis de superposición (fusión de dos o más capas o mapas para identificar áreas de intersección), análisis de red (modelado de las redes y el cálculo de parámetros tales como la distancia más corta entre dos ubicaciones), superficie de área, cálculos de distancia, modelado de superficies tridimensionales (Ward y Carpenter, 2000) y kriging (método de interpolación lineal que predice los valores de una variable en lugares no incluidos en la muestra sobre la base de observaciones en lugares conocidos utilizando un modelo de covarianza de una función aleatoria (Berke, 2004)).

Los avances en la tecnología informática, el fomento de su uso por el gobierno y la amplia disponibilidad de cursos académicos y comerciales sobre los SIG son responsables de su crecimiento.

Algunos ven a los SIG sólo como una herramienta para la investigación y análisis de la política espacial, mientras que otros creen que es parte de una nueva ciencia emergente más grande incluyendo la geografía, la cartografía, la geodesia y la teledetección (Mullner, Chung, Croke y Mensah, 2004).

El mayor potencial de los SIG es su capacidad para mostrar con claridad los resultados de los análisis complejos a través de mapas.

Los problemas en el uso de los SIG incluyen sus costos, la necesidad de capacitar adecuadamente al personal, el uso de unidades espaciales adecuadas, y el riesgo que supone violar la confidencialidad del paciente.

En la actualidad existe un movimiento hacia la relación con los SIG como una ciencia (Ciencias de Información Geográfica) en lugar de simplemente una tecnología (Goodchild, 2000; Rinaldi et al., 2006).

### **3.2. Sensado Remoto**

Se utilizó el término "sensado remoto" (SR) por primera vez - en los EE.UU. durante la década de 1960 para designar la técnica que permite el estudio de los objetos sin ningún contacto directo, a través de la captura de imágenes (Rinaldi et al., 2006).

SR es la ciencia de adquirir información sobre la superficie de la Tierra sin estar realmente en contacto con ella. Esto se realiza mediante la detección y grabación de energía reflejada o emitida y su procesamiento, análisis y aplicación (Fundamentals of

Remote Sensing - Introduction, 2015). El portador de esta información es la radiación electromagnética. El principal emisor es el sol, su radiación es reflejada por la tierra y los objetos y puede ser capturado por sensores pasivos a bordo de satélites. Otra opción es que el propio sensor incorpore un emisor de radiación (sensores activos) cuyo reflejo es capturado posteriormente por el mismo sensor (Bustos, Frery, Lamfri y Scavuzzo, 2004).

La resolución temporal, espacial y espectral son factores muy importantes para determinar la utilidad de un sensor satelital para estudios epidemiológicos y de salud pública (Tatem, Goetz y Hay, 2004). El término resolución espacial designa el objeto más pequeño que puede distinguirse sobre una imagen, en un sistema fotográfico, se mide como la separación mínima en la que los objetos aparecen distintos y separados en la fotografía y en sistemas óptico-electrónicos está determinado por el tamaño del píxel (la unidad más pequeña y homogénea de una imagen digital), medido en metros por encima de la tierra (Chuvienco, 2008). La resolución temporal, se refiere al tiempo de revisita, es decir la frecuencia con la que el sensor toma imágenes de la misma parte de la superficie de la tierra (Chuvienco, 2008). Por otra parte, la resolución espectral indica el número y la anchura de las bandas espectrales (rango del espectro electromagnético, en el que un sensor captura la radiación) que puede discriminar el sensor. Cuanto mayor es la resolución espectral, un sensor es más adecuado para caracterizar espectralmente diferente cobertura de la tierra (Chuvienco, 2008). Por último, la resolución radiométrica es la capacidad del sensor para detectar variaciones en la radiancia espectral recibida. Para los sistemas fotográficos se indica por los niveles de gris de la película y para las imágenes digitales es indicado por el rango de valores expresado como el número de bits de la imagen que se almacena (Bustos, 2004; Chuvienco, 2008). Los cuatro aspectos de la resolución están estrechamente relacionados y en general cuando aumenta la resolución espacial disminuye la resolución temporal y espectral (Figura 2, apéndice A).

Algunos satélites pasivos (por ejemplo: Landsat, SPOT, IKONOS) y activos (por ejemplo: TerraSAR-X, RADARSAT-2, Cosmo Skymed) proporcionan imágenes de alta y muy alta resolución espacial (de unos 30 a 1 metro), pero a expensas de la resolución temporal (que puede ser de varios días o semanas). Otras plataformas como METEOSAT, GEOS, AQUA\_TERRA / MODIS, NOAA AVHRR, pueden cubrir el mismo lugar una o más veces al día, siendo adecuados para estudios a gran escala (nivel de continente o país), pero no es adecuado para el análisis local (Machault et al., 2012).

Las variables que requieren alta resolución espacial (uso y cobertura del suelo) se pueden estudiar con sensores de este tipo y las que varían en el tiempo (precipitaciones, temperatura y humedad del aire) puede requerir sensores con alta resolución temporal (Kalluri, Gilruth, Rogers, y Szczur, 2007; Machault et al., 2012).

Generalmente imágenes de alta resolución espacial tiene un alto costo y las de baja resolución espacial son libres, (De La Rocque, Michel, Plazanet y Pin, 2004; Kalluri et al., 2007; Machault et al., 2012). Dependiendo del número y ancho de banda suele distinguirse entre sistemas multispectrales e hiperespectrales.

Ejemplo de satélites multispectrales son: Landsat y SPOT, que se caracterizan por un número de bandas espectrales no muy altas (por ejemplo: LANDSAT\_8 tiene 9 bandas). Sistemas hiperespectrales (espectrómetros de imagen) se caracterizan por la grabación de imágenes en cientos de bandas espectrales muy estrechas (por ejemplo: HYPERION EO- sensor 1 de satélites de la NASA, tiene 220 bandas).

En resumen, la elección del sensor depende de los objetivos y el nivel de precisión requerido para el estudio, la escala de trabajo (escala global: sensores de menor resolución espacial, escala regional: sensores de mayor resolución espacial), la dinámica del fenómeno estudiado (fenómenos dinámicos requieran sensores de mayor resolución temporal), espectro en que el fenómeno de interés se manifiesta (sensores con bandas espectrales que nos ofrezcan información del fenómeno estudiado) y no menos importante, el costo de las imágenes satelitales.

La información derivada de SR y otras formas de datos espaciales puede usarse para soportar una amplia gama de análisis epidemiológicos gracias al potencial de proporcionar una cobertura completa en la forma de una o más imágenes, abordando a una variedad de dominios espaciales y temporales. Ninguna otra fuente de información ambiental puede proporcionar esta valiosa información espacial. Esto es especialmente cierto a escala global (Atkinson y Graham, 2006). Técnicas de SR se pueden utilizar para identificar y monitorear factores que intervienen en la transmisión de enfermedades (De La Rocque et al., 2004). La Interpretación visual y el análisis multi-temporal y multi-espectral, de productos de datos derivados de satélites de observación de la Tierra, se utilizan para definir y mapear el hábitat y la ecología de vectores de enfermedades (Palaniyandi, 2012).

Como se mencionó previamente en la introducción el NDVI, cobertura y uso del suelo, elevación, pendiente, aspecto, capas geológicas, presencias de cuerpos de agua, temperatura, precipitaciones y humedad son comúnmente usadas para fines de investigación en epidemiología.

## **TEMPERATURA**

La temperatura superficial de la tierra (LST, por sus siglas en ingles), de día y de noche, puede medirse a través del sensor MODIS (Moderate Resolution Imaging Spectroradiometer, <http://modis-land.gsfc.nasa.gov/temp.html>) a bordo de los satélites Terra y Aqua y otros sensores infrarrojos térmicos (METEOSAT, GOES y NOAA-VHRR).

LST se correlaciona bien con la temperatura que prevalece en el aire, pero la cobertura del suelo, la humedad, las condiciones atmosféricas y el período del día puede introducir aberraciones (Colombi, Pepe, Rampini y Michele, 2007; Vancutsem, Ceccato, Dinku y Connor, 2010). Sin embargo, un estudio en África comparando datos LST provenientes de sensores remotos con datos interpolados de estaciones meteorológicas en tierra, llegó a la conclusión de que LST media anual, corresponde con la temperatura del aire dentro de un rango de +/- 4 ° C (Hay y Lennon, 1999). Otro artículo informa de una correlación significativa entre las observaciones LST y temperatura de estaciones terrestres, tanto en África como en Europa, con la raíz del error cuadrático medio de alrededor de 2° C. Estudios realizados a partir de LST MODIS con el fin de estimar la temperatura superficial del aire, fueron realizados en México encontrando correlaciones significativas a nivel nacional (Mas, 2011) y en una región de Chile, estimando la temperatura del aire a una altura de 2m a partir de datos obtenidos por teledetección con un error de 3.72 °C (Flores y Lillo, 2010). Por lo tanto, los datos LST son útiles, pero conviene sacar conclusiones con cuidado (Machault et al., 2011).

## VEGETACIÓN

La vegetación también puede detectarse, puede determinarse a través del NDVI, el más utilizado en salud (Machault et al., 2011). NDVI incluye indirectamente el efecto de la temperatura, la humedad, la lluvia, la altitud del sol y uso y cobertura de la tierra (Britch et al., 2008). Se puede utilizar como sustituto de la precipitación en ciertos períodos y áreas (Justice CO et al., 1991). El NDVI varía entre -1 y +1, basado en la óptima absorción de la vegetación en las longitudes de onda rojas (R) y reflexión máxima del infrarrojo cercano (NIR), esto es  $NDVI = (NIR - R) / (NIR + R)$ .

El NDVI no sólo depende de la cobertura de la tierra, sino también de las condiciones atmosféricas, no existe un umbral estricto para la detección de la vegetación. Sin embargo, un valor superior a 0,2 corresponde generalmente a un área cubierta por vegetación y valores negativos corresponden a agua, infraestructuras generales o asfalto. Varios instrumentos de baja resolución, por ejemplo, MODIS y AVHRR, proporcionan valores de NDVI, pero también se pueden calcular mediante la combinación de las bandas espectrales apropiadas de cualquier imagen de alta resolución.

## PRECIPITACIÓN

Se puede medir en forma directa o por métodos indirectos. La duración de nubes frías se puede derivar desde imágenes térmicas Metaosat. Los satélites de la Misión de Medición de Lluvias Tropicales (TRMM; <http://trmm.gsfc.nasa.gov/>) proporcionan casi en tiempo real, la medición directa de las lluvias en las zonas tropicales a base de microondas pasivas y sensores activos (radar), que coincide estrechamente con observaciones en tierra de pluviómetros (Sharma, Samarkoon, Hazarika y Kafle, 2007), incluso en las zonas urbanas (Han, Burian y Shepherd, 2011). Sin embargo, la influencia de la ubicación, el clima, la topografía, período del tiempo, tipos de nubes y tipos de precipitaciones puede afectar la precisión.

### 3.3. Epidemiología Espacial y Análisis Ecológico

La epidemiología espacial se ocupa de la descripción y el análisis de la variación geográfica o espacial de la enfermedad con respecto a factores de riesgo demográficos, ambientales, socio-económicos, genéticos e infecciosos (Elliott y Wartenberg, 2004). El análisis epidemiológico espacial tiene como objeto la descripción de los patrones espaciales, la identificación de grupos de enfermedades y la explicación o predicción del riesgo de enfermedad (Pfeiffer et al., 2008).

Por otra parte la descripción de las relaciones existentes entre la distribución geográfica de las enfermedades y los factores de riesgo ambientales y su análisis por medio de procedimientos estadísticos se denomina Análisis Ecológico (Kistemann et al., 2002). Un gran número de estudios se han llevado adelante con el fin de analizar la relación entre indicadores de la enfermedad (ej. positivos, incidencia y prevalencia) y variables

explicativas ambientales y climáticas (Herbreteau, Salem, Souris, Hugot y Gonzalez, 2005).

Como se dijo previamente, SIG y SR ofrecen medios eficaces para el mapeo de la enfermedad, el análisis ecológico y la vigilancia epidemiológica, y se han convertido en herramientas indispensables para el procesamiento, análisis y visualización de datos espaciales (Rinaldi et al., 2006).

## CAPÍTULO 4:

### 4. MAPEO DE ENFERMEDADES Y MODELOS JERÁRQUICOS BAYESIANOS

#### 4.1. Mapeo de enfermedades Bayesiano

Los estudios sobre mapeo de Enfermedades pretenden resumir la variación espacial del riesgo de enfermar, con el fin de evaluar y cuantificar la verdadera cantidad de heterogeneidad espacial y los patrones asociados, para destacar áreas de riesgo elevado o bajo y obtener pistas sobre la etiología de la enfermedad.

El desarrollo de métodos de mapeo de la ocurrencia de la enfermedad ha progresado considerablemente en los últimos años. Uno de los ejemplos más tempranos de cartografía de la enfermedad es el mapa de las direcciones de víctimas del cólera relacionados con la ubicación de las fuentes de agua realizado por Snow (1854). En ese caso, se registraron y evaluaron los domicilios de las víctimas y su proximidad a las fuentes de contaminación putativas (bombas de suministro de agua). La representación y análisis de mapas de datos de ocurrencia de enfermedad ahora se establece como una herramienta básica en el análisis regional en salud pública.

El interés creciente en cartografía de enfermedades ha dado lugar a un mayor uso de herramientas estadísticas geográficas o espaciales y al análisis de datos recogidos de manera sistemática tanto para fines de salud pública como de estudios ecológicos en relación a variables explicativas.

El estudio de la distribución geográfica de la enfermedad puede tener una variedad de usos. Las principales áreas de aplicación pueden ser desglosadas en, cartografía de la enfermedad, análisis clúster y análisis ecológico (Lawson, Browne y Rodeiro, 2003). En la primera clase, por lo general el objeto del análisis es proporcionar (estimar) el verdadero riesgo relativo de una enfermedad de interés a través de un área geográfica de estudio. Estos métodos son útiles entre otras cosas, para la asignación de recursos en servicios de salud y en la construcción de atlas de enfermedad. La segunda clase, análisis clúster, tiene especial importancia en la vigilancia epidemiológica, ya que es importante, evaluar si una enfermedad está agrupada en el espacio y donde se encuentran los conglomerados. Esto puede conducir a un examen de los posibles peligros ambientales. La tercera clase, el análisis ecológico del que nos explayaremos

luego, es de gran relevancia dentro de la investigación epidemiológica, su enfoque es el análisis de la distribución geográfica de la enfermedad en relación con covariables explicativas, por lo general a un nivel espacial agregado.

Se tienen en cuenta dos situaciones que surgen con frecuencia en los estudios de la distribución geográfica de la enfermedad. Estas situaciones se definen por la forma de recolección de los datos planteada en tales estudios. En primer lugar, un área de estudio o ventana se define y dentro de esta se registran las ubicaciones de casos de una enfermedad específica por un período fijo de tiempo. Estos lugares suelen ser direcciones residenciales (dirección de la calle, o en una escala espacial más alta, códigos postales). Cuando se conocen tales direcciones es posible proceder por análisis directo de los lugares de los casos. Esto se denomina análisis “caso-evento” (Lawson et al., 2001; Elliot, Wakefield, Best, y Briggs, 2000), para un ejemplo en WinBUGS ver Congdon (2014). A menudo, este análisis requiere el uso de modelos de procesos puntuales y metodología asociada (Lawson, 1989; Baddeley y Turner 2000; Diggle, 2003; Illian, Penttinen, Stoyan y Stoyan, 2008). Debido a los requisitos de confidencialidad médica, a menudo no es posible obtener datos a este nivel de resolución y se debe recurrir al análisis del recuento de eventos de enfermedad dentro de un área definida dentro de la ventana de estudio (Lawson et al., 2001; Pfeiffer et al., 2008). Estas pequeñas áreas son regiones arbitrarias generalmente definidas a efectos administrativos, como las secciones censales, municipios, departamentos, distritos o regiones sanitarias de un distrito o provincia.

Estimación de pequeñas áreas es una técnica estadística que implica la estimación de parámetros para pequeñas subpoblaciones (áreas pequeñas) cuando una muestra es insuficiente o nula para la subpoblación (área pequeña) a fin de poder hacer estimaciones precisas sobre ellas (Rao, 2003; Clement, 2014). El término "área pequeña" puede referirse estrictamente a una zona geográfica pequeña, como un condado, pero también puede referirse a un "pequeño dominio", es decir, un grupo demográfico particular dentro de un área (Rao, 2003; Clement, 2014). Métodos de estimación para áreas pequeñas utilizan modelos y fuentes de datos adicionales (como los datos del censo) que existen para estas pequeñas áreas con el fin de mejorar las estimaciones.

Los datos de este tipo consisten en cuentas de casos dentro de extensiones y el análisis de estos datos se denomina análisis de “conteos agrupados por áreas”. Nosotros nos centramos exclusivamente en éstos.

En esencia, el recuento es una agregación de todos los casos dentro de cada área. Debido a la agregación, las referencias espaciales de los casos individuales (ubicaciones) se pierden y por lo tanto cualquier georreferencia del conteo está relacionada con la ubicación de las áreas. En una ventana de estudio elegida se encuentran  $m$  áreas. Denotemos los conteos de enfermedad dentro de  $m$  áreas como  $\{y_i\}, i = 1, \dots, m$ .

#### 4.1.1. REPRESENTACIÓN BRUTA DE LA DISTRIBUCIÓN DE LA ENFERMEDAD

La forma más sencilla posible de cartografía de enfermedades, es la representación de las tasas de enfermedad en grupos específicos de ubicaciones. Para recuentos dentro de áreas, esta es una representación gráfica del número de eventos en áreas en un conjunto adecuado de ubicaciones. Las ubicaciones de los casos de enfermedad dentro de una población espacialmente heterogénea pueden mostrar información sobre el patrón general dentro de una ventana. Sin embargo, cualquier interpretación de la estructura de estos eventos está severamente limitada por la falta de información sobre la distribución espacial de la población de referencia "en riesgo de enfermar" y que dio lugar a los casos (Lawson et al., 2001). Esta población también tiene una distribución espacial y el no tener en cuenta esta variación limita severamente la capacidad de interpretar el mapa de casos-eventos resultante. En esencia, las zonas de alta densidad de población "en riesgo de enfermar" tenderían a producir mayor cantidad casos-eventos y así, sin tener en cuenta esta distribución, áreas de alta intensidad de la enfermedad podrían ser falsamente atribuidos a un exceso de riesgo de enfermedad.

En el caso de recuentos de casos de enfermedad por áreas, consideraciones similares se aplican cuando se construyen mapas de recuento bruto. Aquí, la variación en la densidad de la población afecta también a la incidencia espacial de la enfermedad (Lawson et al., 2001). En general son preferibles, los métodos que tratan de incorporar la población.

#### 4.1.2. RAZÓN DE MORTALIDAD/MORBILIDAD ESTANDARIZADA

Para evaluar el estado de un área con respecto a la incidencia de la enfermedad, es conveniente tratar de evaluar primero, qué incidencia de enfermedad se "espera" localmente en cada área de la zona de estudio y luego comparar la incidencia observada con la incidencia "esperada" (Lawson et al., 2003). Este enfoque ha sido utilizado tradicionalmente para el análisis de recuentos dentro de áreas. Tradicionalmente, la relación de los recuentos observados y esperados dentro de áreas se llama Razón de Mortalidad/Morbilidad Estandarizada o "standardized morbidity rate" (SMR) por sus siglas en inglés y esta relación es una estimación del riesgo relativo dentro de cada área (la relación describe las probabilidades de estar en el grupo de enfermos en lugar de la población de fondo, o lo que es lo mismo, la población de donde se originan los eventos).

Definimos  $y_i$  como el conteo observado de casos de enfermedad en el área  $i$ , y  $e_i$  es el conteo esperado dentro del área  $i$ :

$$E_i = n_i \left[ \frac{\sum_{i=1}^N O_i}{\sum_{i=1}^N n_i} \right] \quad (1)$$

donde  $n_i$ , es la población observada en riesgo en el área  $i$ .

Entonces la SMR es definida como:

$$\hat{\theta}_i = \frac{y_i}{e_i} \quad (2)$$

donde el término  $\hat{\theta}_i$  es la SMR para cada área  $i$ .

En este caso se debe decidir si se expresa  $\hat{\theta}_i$  como patrón de relleno en cada región, o localizando el resultado en algún lugar específico, como el centro de gravedad del área. Si se decide que estas medidas deben ser consideradas como regiones continuas a través de las áreas, entonces puede hacerse la interpolación de  $\hat{\theta}_i$ .

SMR se utiliza comúnmente en el mapeo de enfermedad, pero tiene muchos inconvenientes. Los mapas de SMR tienden a ser dominados por zonas con un pequeño número de observaciones, o si se centran en la significación estadística de las estimaciones, tienden a estar dominados por áreas con grandes poblaciones, pero potencialmente menos importantes en la aparición de enfermedad (Leyland y Davies, 2005). Además, SMR se basa en estimadores de razón y por lo tanto puede producir grandes cambios en las estimaciones con cambios relativamente pequeños en el valor esperado (denominador). En el caso de que los esperados tienden a cero el SMR será muy grande para cualquier cuenta positiva. También los SMR con valor cero no distingue la variación en los recuentos esperados y la varianza de SMR es proporcional a  $1/e_i$ . SMR es esencialmente una estimación saturada de riesgo relativo y por lo tanto no es parsimoniosa.

## 4.2. Modelos jerárquicos Bayesianos

En el punto anterior se han destacado algunos métodos sencillos para la estimación del riesgo relativo y se ha insistido en la deficiencia de la SMR como un estimador de riesgo relativo. Para sortear estas deficiencias se han propuesto una variedad de estimadores. Un enfoque para mejorar el proceso de estimación es controlar el comportamiento de  $\{\theta_i\}$ . Para este fin, es conveniente ampliar el modelo suponiendo que los parámetros tienen distribuciones. Estas distribuciones controlan la forma de los parámetros y se especifican por el investigador con base, por lo general, en su creencia previa sobre su conducta. Estas distribuciones se denominan “distribuciones a priori”. Vamos a denotar una distribución a priori como  $g(\theta)$ . En el contexto de cartografía de

enfermedades, una distribución a priori comúnmente asumida por  $\theta$  es la distribución gamma y el modelo resultante es el modelo Poisson-gamma (Lawson et al., 2003).

#### 4.2.1. VEROSIMILITUD Y DISTRIBUCIÓN A POSTERIORI

Distribución a priori y verosimilitud proporcionan dos fuentes de información acerca de cualquier problema. La verosimilitud informa sobre el parámetro a través de los datos, mientras que la distribución a priori informa a través de las creencias o suposiciones previas. Cuando hay grandes cantidades de datos, es decir, el tamaño de la muestra es grande, la verosimilitud contribuirá más a la estimación del riesgo relativo. Cuando los datos son pocos la distribución a priori domina el análisis.

Note que la verosimilitud del dato  $\{y_i\}$  dado los parámetros  $\{\theta_i\}$  como  $L(y|\theta)$  y el log-verosimilitud como  $l(y|\theta)$ . Note que  $\theta$  no tiene que ser de la misma dimensión que  $y$ . El producto de la verosimilitud y la distribución a priori es llamado “distribución a posteriori”. Esta distribución describe el comportamiento de los parámetros luego que los datos son observados y las asunciones previas son hechas. La distribución a posteriori es definida como:

$$p(\theta|y) \propto L(y|\theta)g(\theta)$$

donde  $g(\theta)$  es la distribución a priori conjunta del vector  $\theta$ .

Un ejemplo simple de este tipo de modelo en mapeo de enfermedades ocurre cuando la verosimilitud del dato es Poisson y hay un riesgo relativo con distribución a priori del tipo gamma:

$$p(\theta|y) \propto L(y|\theta)g(\theta)$$

donde  $g(\theta)$  es una distribución gamma con parámetros  $\alpha, \beta$ , ej: Gamma ( $\alpha, \beta$ ) y  $L(y|\theta) = \prod_{i=1}^m \{(e_i\theta)^{y_i} \exp(-e_i\theta)\}$  barra una constante solo dependiente del dato. Una notación compacta del modelo es:

$$y_i|\theta \sim \text{Poisson}(e_i\theta),$$

$$\theta \sim \text{Gamma}(\alpha, \beta).$$

#### 4.2.2. MODELOS JERÁRQUICOS

Anteriormente se dieron ejemplos simples de verosimilitud y distribución a priori. En este ejemplo la distribución a priori para el parámetro también tiene parámetros que controlando su forma. Estos parámetros  $(\alpha, \beta)$  puede asumir valores, pero generalmente un investigador no tendrá una fuerte creencia previa en los valores de los parámetros.

Por lo tanto, el investigador puede querer estimar estos parámetros a partir de los datos. Alternativamente y de manera más formal, los parámetros dentro de los modelos son considerados como estocásticos (y por lo tanto tienen distribuciones de probabilidad que gobiernan su comportamiento), entonces, estos parámetros deben también tener distribuciones. Estas distribuciones se conocen como “distribuciones hiperprior” y los parámetros se conocen como hiperparámetros.

La idea de que los valores de los parámetros podrían surgir a partir de distribuciones es una característica fundamental de la metodología bayesiana y conduce naturalmente a la utilización de modelos en los que se presentan los parámetros dentro de jerarquías. En el ejemplo del modelo Poisson-gamma existe una jerarquía de dos niveles:  $\theta$  tiene una distribución  $Gamma(\alpha, \beta)$  en el primer nivel de la jerarquía y  $\alpha$  y  $\beta$  tendrán una distribución hyperprior  $(h_\alpha)$  y  $(h_\beta)$  respectivamente, en el segundo nivel de jerarquía. Esto se puede escribir como:

$$y_i | \theta \sim Poisson(e_i \theta),$$

$$\theta | \alpha, \beta \sim Gamma(\alpha, \beta),$$

$$\alpha | v \sim h_\alpha(v).$$

$$\beta | p \sim h_\beta(p).$$

Claramente es importante terminar la jerarquía en un lugar adecuado, de lo contrario se podría suponer una jerarquía infinita de los parámetros. Por lo general, el punto de corte se elige para estar donde una futura variación en los parámetros no afectará el modelo a nivel más bajo. En este punto los parámetros se asumen como fijos. Por ejemplo, en el modelo de Poisson-gamma, si se asume que  $\alpha$  y  $\beta$  son fijos entonces la distribución a priori gamma será fija y se elegirán  $\alpha$  y  $\beta$  como no informativas, los datos no informan acerca de la distribución en absoluto. Sin embargo, se puede permitir un mayor nivel de variación, por ejemplo, hiperprior para  $\alpha$  y  $\beta$ , donde se fijan los valores de  $v$  y  $p$  sin influir en gran medida la variación del nivel inferior.

#### 4.2.3. INFERENCIA A POSTERIORI

Cuando se emplea un modelo de probabilidad simple, a menudo se utiliza máxima verosimilitud para proporcionar una estimación puntual y de variabilidad asociada para

los parámetros. Esto es cierto para los modelos simples de mapeo de enfermedades. Por ejemplo en el modelo  $y_i|\theta \sim Poisson(e_i\theta)$  la máxima verosimilitud estimada para  $\theta$  es la tasa global de la región de estudio, ej.  $\sum y_i / \sum e_i$ . Por otro lado, la SMR es la estimación de máxima verosimilitud para el modelo  $y_i|\theta \sim Poisson(e_i\theta)$ .

Cuando se emplea un modelo jerárquico Bayesiano ya no es posible proporcionar una estimación puntual simple para cualquiera de los  $\theta_i$ . Esto es porque se supone que ya no se fija el parámetro, si no que surge de una distribución de valores posibles. Dados los datos observados, el parámetro o parámetros de interés se describirán por la distribución a posteriori, y por lo tanto esta distribución se debe encontrar y examinar. Es posible examinar el valor esperado (media) o el modo de la distribución a posteriori para dar una estimación puntual de un parámetro. Así como la estimación de máxima verosimilitud es el modo de la probabilidad, entonces el máximo una estimación a posteriori es aquel valor del parámetro o parámetros en el modo de la distribución a posteriori. Más comúnmente se utiliza el valor esperado del parámetro o parámetros. Esto se conoce como la media a posteriori (o estimación de Bayes). Para distribuciones simétricas unimodales simples, el modo y la estimación de la media coinciden.

Para algunas distribuciones a posteriori simples, es posible encontrar la forma exacta de la distribución a posteriori y encontrar formas explícitas para la media posterior o el modo. Sin embargo, es común, en el caso de los modelos razonablemente realistas dentro de la cartografía de enfermedades, que no sea posible obtener una forma cerrada para la distribución a posteriori. Por lo tanto, a menudo no es posible derivar estimadores simples para parámetros tales como el riesgo relativo. En esta situación compleja se debe hacer el muestreo a posteriori, es decir, utilizando métodos de simulación para obtener muestras de la distribución a posteriori que luego puedan resumirse para producir estimaciones de las cantidades pertinentes.

#### 4.2.4. MÉTODOS MARKOV CHAIN MONTE CARLO

A menudo en cartografía de enfermedades, los modelos realistas tienen dos o más niveles y la complejidad resultante de la distribución a posteriori de los parámetros requiere el uso de algoritmos de muestreo. En adición los modelos flexibles requieren conmutación entre una variedad de modelos relativamente complejos. En este caso es conveniente tener un método de muestreo a posteriori flexible para ser aplicado a una variedad de modelos.

Los métodos de cadenas de Markov Monte Carlo (MCMC) son métodos de simulación para extraer la distribución a posteriori  $p(\theta|y)$  dado un modelo, una verosimilitud  $L(y|\theta)$  y un dato  $y$ , utilizando secuencias dependientes de variables aleatorias. Es decir, MCMC produce una muestra de la distribución a posteriori de un parámetro, de la que se pueden estimar cantidades de interés a posteriori.

MCMC han existido desde hace casi tanto tiempo como las técnicas de Monte Carlo, a pesar de que su impacto sobre la estadística no se ha sentido verdaderamente hasta principios de 1990, excepto en los campos especializados de Estadística Espacial y Análisis de Imágenes, donde esos métodos aparecieron antes (Robert y Casella, 2011).

Ahora el muestreo posterior es lugar común en una variedad de software que han incorporado estos métodos (incluyendo: WinBUGS, OpenBUGS y JAGS). El software R puede llamar funciones para “inferencia bayesiana usando muestreo de Gibbs”, BUGS por sus siglas en inglés. Para mayor información sobre métodos MCMC y aplicaciones reales puede consultarse literatura como (Gilks, Richardson, y Spiegelhalter, 1996).

En los métodos MCMC se simulan valores de una densidad propuesta sucesivamente, que no tiene que ser necesariamente parecida a la densidad a posteriori. Cada valor generado depende sólo del anterior valor simulado, de ahí la noción de cadena de Markov. Estos métodos son muy usados en la computación moderna Bayesiana ya que en la mayoría de los casos la forma analítica de  $p(\theta|y)$  es desconocida y la dimensión de  $\theta$  elevada.

## MUESTREO DE GIBBS

MCMC fue desarrollado en 1953 por los físicos Metropolis et al., y posteriormente generalizadas por Hastings (1970), por lo que uno de los principales algoritmos MCMC se denomina Metropolis Hastings. Muchas técnicas diferentes de MCMC están disponibles ahora. Una de las técnicas MCMC más utilizadas es el muestreo de Gibbs (Geman y Geman, 1984), se basa en la idea de que, para resolver un gran problema, en lugar de tratar de hacer todo a la vez, es más eficaz romper el problema en subunidades más pequeñas y resolver una a la vez.

Aquí está un bosquejo de cómo funciona el muestreo de Gibbs (Kéry, 2010). Siendo nuestros datos  $y$ , y nuestro vector desconocido  $\theta$  consistentes en  $k$  subcomponentes  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ , por lo tanto, queremos estimar  $k$  parámetros.

1. Elegir los valores iniciales  $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}$
2. Muestra  $\theta_1^{(1)}$   $p(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, y)$
- Muestra  $\theta_2^{(1)}$   $p(\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, y)$
- .....
- Muestra  $\theta_k^{(1)}$   $p(\theta_k | \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{k-1}^{(1)}, y)$
3. Repetir el paso 2 muchas veces (ejemplo: 100, 1000) eventualmente obteniendo una muestra de  $p(\theta|y)$ .

El paso 2 es llamado actualización o iteración del muestreo de Gibbs y luego de que la convergencia es alcanzada, esta conduce a una extracción (= muestra) consistente en  $k$  valores desde la distribución a posteriori conjunta  $p(\theta|y)$ . La distribución condicional en este paso es llamada “condicional completa” porque ella condiciona a todos los parámetros. La secuencia de sorteos al azar para cada uno de los parámetros  $k$  resultantes de la etapa 3 forma una cadena de Markov.

### 4.3. Modelo de Poisson para conteo de datos.

Cuando es de interés una variable respuesta, que es un recuento del número de eventos que ocurren en una población de un determinado tamaño, o un recuento del número de eventos en relación con el número de personas o animales por unidad de tiempo en riesgo, una suposición razonable es que estos conteos sigan una distribución de Poisson, especialmente cuando una enfermedad es no contagiosa y/o rara (Pfeiffer et al., 2008). Quizás el modelo más utilizado para datos de conteo en áreas pequeñas es el modelo de Poisson. Este modelo es apropiado cuando hay bajo conteo de enfermedad y la población es relativamente grande en cada área (Lawson, 2009).

A menudo el conteo de enfermedad  $y_i$  asume una media  $u_i$  y se distribuye independientemente como

$$y_i \sim \text{Poisson}(u_i)$$

La verosimilitud está dada por

$$L(y|u) = \prod_{i=1}^m u_i^{y_i} \exp(-u_i) / y_i!$$

La función de la media generalmente consiste en dos componentes, el primer componente representa el efecto de la población de fondo y el segundo componente representa el exceso de riesgo dentro de un área, a menudo denominado *Riesgo Relativo*. El primer componente es estimado o computado por comparación de tasas de la enfermedad en una población estándar, obteniendo una tasa local esperada. Esto se denomina a menudo estandarización (Inskip et al., 1983). Por lo tanto, se asume generalmente que el dato está distribuido independientemente con expectativa.

$$E(y_i) = u_i = e_i \theta_i$$

Donde  $e_i$  es la tasa esperada para el área  $i$  y  $\theta_i$  es el riesgo relativo para el área  $i$ . El conteo de la enfermedad observado en cada área  $i$  asume una distribución de Poisson con media  $u_i$  y es función de la cantidad de enfermedad esperada (que depende del tamaño de la población) y el término de riesgo relativo en un área específica. Como se desarrollan modelos jerárquicos Bayesianos, se considera que  $\{y_i\}$  es condicionalmente independiente, dado el conocimiento de  $\{\theta_i\}$ . La tasa esperada generalmente se asume fija para el periodo de tiempo considerado en el ejemplo espacial, aunque la literatura de la estimación de pequeñas áreas sugiere que esto puede ser ingenuo (Ghosh y Rao, 1994; Rao, 2003).

Usualmente el foco de interés es el modelado del riesgo relativo. La aproximación común a esto es asumir un enlace logarítmico a un modelo de predicción lineal:

$$\log \theta_i = n_i$$

Esta forma de modelo ha tenido un uso generalizado en el análisis de datos de recuento en áreas pequeñas en una gama de aplicaciones (Stevenson et al., 2005; Waller y Gotway, 2004).

#### ESPECIFICACIÓN DEL PREDICTOR LINEAL

En el modelo previo se especifica una función predictora ( $n_i$ ) para relacionar la media de la variable respuesta aleatoria, a través de una función de enlace adecuada. A menudo la función predictora se asume lineal como función de covariables (o variables explicativas) fijas y posiblemente de efectos aleatorios (Lawson, 2009). Se define esto en una forma general:

$$n_i = x_i' \beta$$

para  $m$  covariables, como

$$n_i = \alpha + \beta x_i + \dots + \beta_m x_m$$

Siendo el riesgo relativo función de una serie de  $x_{mi}$  variables explicativas o covariables más el error para cada región.

$$\log(\theta_i) = (\alpha + \beta x_i + \dots + \beta_m x_m) + \varepsilon_i$$

De esta manera una estimación del número real de conteos de enfermedad en cada área viene dado por:

$$\log(u_i) = \log(e_i) + (\alpha + \beta x_i + \dots + \beta_p x_p) + \varepsilon_i$$

El término  $\varepsilon_i$  representa los residuos del riesgo relativo en el área  $i$  después de ajustar por covariables incluidas en el modelo. A menudo es conveniente interpretar  $\varepsilon_i$  como residuos que reflejan la variabilidad entre áreas, debido a factores de riesgo desconocidos o no medidos (Pfeiffer et al., 2008).

El  $\log(e_i)$  es un factor de ajuste conocido como offset y representa el ajuste de acuerdo con áreas de diferentes tamaños de población (si todas las áreas tienen igual tamaño de población no se requiere el término offset), (Pfeiffer et al., 2008).

Para los datos de conteo en áreas pequeñas, es probable que las covariables se obtengan en la escala de la zona (Lawson, 2009). Por ejemplo, para las secciones censales, es probable que sean variables socioeconómicas (ej. nivel de ingreso, pobreza, etc.) disponibles a partir de censos. Además, algunas variables aleatorias podrían ser incluidas como variables supra-zona como el distrito de salud en el que el área se encuentra. Covariables ambientales también pueden ser interpoladas para la utilización en el nivel de la sección censal. Por ejemplo, medidas de contaminación del aire podrían ser promediadas a lo largo del tracto.

## MODELO LINEAL BAYESIANO

En el paradigma Bayesiano todos los parámetros son estocásticos. Por lo tanto, en el modelo de la covariable

$$n_i = x_i' \beta,$$

se asume que el parámetro  $\beta$  tiene distribución a priori. Por lo tanto, esto se puede formular como

$$P(\beta, \tau_\beta | data) \propto L(data | \beta, \tau_\beta) f(\beta, \tau_\beta)$$

Donde  $f(\beta, \tau_\beta)$  es la distribución conjunta de la covariable condicional al hiperparámetro  $\tau_\beta$ . A menudo consideramos estos parámetros como independientes y entonces

$$f(\beta, \tau_\beta) = \prod_{j=1}^p f_j(\beta_j, \tau_{\beta_j})$$

Generalmente se asume que el parámetro “covariable” puede ser descrito por una distribución Gaussiana y si se permite que los parámetros se correlacionen entonces se podría tener la especificación gaussiana multivariante:

$$f(\beta, \tau_\beta) = N_p(0, \Sigma_\beta)$$

Donde bajo esta asunción de prior,  $E(\beta, \tau_\beta) = 0$  y  $\Sigma_\beta$  es la covarianza condicional de los parámetros. La especificación común asume prior independiente y es:

$$f(\beta, \tau_\beta) = \prod_{j=1}^p N(0, \tau_{\beta_j}),$$

donde  $N(0, \tau_{\beta_j})$  es una variable simple con distribución Gaussiana, de media cero y varianza  $\tau_{\beta_j}$ . En este punto, se hace normalmente, una suposición acerca de la variación en los hiperparámetros. En el próximo nivel de jerarquía se asumen distribuciones hiperprior para  $\tau_\beta$ . La definición de estas distribuciones puede ser importante para el comportamiento del modelo. Por ejemplo, si se asume una hiperprior vaga (juega un papel mínimo en la distribución a posteriori (Gelman, 2014)) para  $\tau_{\beta_j}$ , esta puede llevar a extravariación, presentándose cuando se limita el aprendizaje disponible desde los datos. Esto puede afectar el cálculo del criterio de información de desviación (DIC) y el

diagnóstico de convergencia. Mientras hiperprior uniformes (en un rango positivo amplio) pueden conducir a distribuciones posteriores impropias (estas no integran a 1), se ha encontrado que una distribución uniforme para la desviación estándar puede ser útil (Gelman, 2006), ej.  $\sqrt{\tau_{\beta_j}} \sim U(0, A)$ , donde  $A$  tiene un valor positivo grande.

Sugerencias alternativas son generalmente en forma de distribuciones gamma o gamma inversa con grandes variaciones.

Por ejemplo, se propuso el uso de gamma (0.2, 0.0001) con expectativa 2000 y varianza 20000000 (Kelsall y Wakefield, 2002) y también se examinan varias especificaciones alternativas que incluyen gamma (0.001, 0.001) (Banerjee et al., 2004). Una especificación común (Thomas, Best y Lunn, Arnold y Spiegelhalter, 2004) es gamma (0.5, 0.0005) que tiene expectativa de 1000 y varianza 2000000.

Aunque estas especificaciones anteriores conducen a desinformación relativa, su uso ha sido criticado por (Gelman, 2006) a favor de distribución a priori uniformes sobre la desviación estándar.

Para resumir la jerarquía de tales modelos de covarianza, puede observarse la Tabla 1 en el apéndice B: Tablas, que muestra un modelo de covarianza bayesiano jerárquico sencillo, con tres covariables ( $x_1, x_2, x_3$ ) y riesgo relativo definido como  $\theta_i = \exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})$  para el dato  $\{y_i, e_i\}$  para  $m$  regiones. Los parámetros de regresión se supone que tienen distribución a priori gaussiana independientes con media cero.

#### 4.4. Sobredispersión

Los modelos desarrollados previamente están en función de covariables fijas. En la formulación de modelos Bayesianos todos los parámetros son estocásticos y entonces la extensión para la adición de efectos aleatorios es relativamente sencilla (Lawson et al., 2001; Lawson, 2009).

El modelo de regresión simple descripto previamente no captura el grado de variación presente en los datos provenientes de conteos. La sobredispersión o correlación espacial debida a factores de confusión no observados, usualmente no es capturada por modelos de covarianza simples y a menudo es conveniente incluir algún término adicional o términos en un modelo que pueda capturar tales efectos.

Inicialmente, la sobredispersión o extra-variación puede ser acomodada por la inclusión de una distribución a priori para el riesgo relativo (como el modelo Poisson gamma) o por extensión del término lineal o predictor no lineal para incluir un efecto aleatorio adicional (modelo log-normal).

#### 4.5. Modelo Log-Normal

Una forma simple que es muy popular en aplicaciones de mapeo de enfermedades es asumir un enlace entre el predictor lineal ( $n_i$ ) y el parámetro de interés ( $\theta_i$ ). Este ofrece un método conveniente para introducir un rango de covariables y efectos aleatorios no observados dentro de una formulación simple. La estructura general de esta formulación es  $n_i = x_i' \beta + v_i' \gamma$ . La forma simple que envuelve la heterogeneidad no correlacionada es

$$n_i = v_{1i}$$

donde  $v_{1i}$  es el efecto aleatorio no correlacionado.

Un ejemplo de la aplicación de este modelo puede verse en la Tabla 2 en el apéndice B, donde se presenta el código para el software OpenBUGS. Aquí se muestra el riesgo relativo posterior estimado bajo el modelo Poisson-log-normal. El modelo de verosimilitud asumido en este ejemplo es Poisson con  $y_i \sim \text{Poisson}(e_i \theta_i)$  y  $\theta_i \propto +\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + v_{1i}$ . Aquí la extra variación es modelada cómo heterogeneidad no correlacionada (UH) con media cero y distribución a priori Gaussiana  $v_{1i} \sim N(0, \tau_{v1})$ .

#### 4.6. Modelos de heterogeneidad correlacionada

Cómo se vio previamente, incorporar una especificación de Gauss tiene muchas ventajas. En primer lugar, un efecto aleatorio que es log-gaussiano se comporta de una manera similar a una variable aleatoria gamma, pero el modelo de Gauss puede incluir una estructura de correlación. Por lo tanto, para el caso donde se sospecha que los efectos aleatorios están correlacionados, entonces es más simple especificar una forma log - gaussiana para cualquier variación adicional presente. La extensión más simple es considerar componentes aditivos que describen diferentes aspectos de la variación que se piensa existe en los datos.

Hay que tener en cuenta que esta no es la única especificación posible de una estructura prior para el modelo de heterogeneidad correlacionada (CH) (Møller, Syversveen y Waagepetersen, 1998).

Hay muchas formas de incorporar dicha heterogeneidad en los modelos. Es importante incluir a menudo una variedad de efectos aleatorios en un modelo. Por ej. ambos CH y UH pueden ser incluidos (ver Tabla 3 en el Apéndice B). Un método flexible para

incluir estos términos es incluir un término log-linear con efectos aleatorios aditivos. Besag, York y Mollié (1991) sugirieron por primera vez, para recuentos en áreas, una parametrización de la tasa de la forma,

$$\exp\{x_i'\beta + u_i + v_i\},$$

Donde  $x_i'\beta$  es una serie de covariables fijas,  $u_i$  y  $v_i$  es la heterogeneidad correlacionada y no correlacionada respectivamente. Estos componentes tienen entonces distribución a priori separadas. A menudo, la especificación del componente correlacionado se considera que tiene o bien una distribución a priori Gaussiana intrínseca (CAR) o una distribución a priori normal Multivariante completamente especificada.

#### **4.7. Modelo Poisson para conteos de datos con ambos efectos aleatorios (UH y CH).**

Generalmente se persiguen dos objetivos cuando se modelan datos de área. El primero es cuantificar la influencia de EF en la enfermedad para cada área o región de estudio y el segundo, identificar áreas donde hay conteos de enfermedad elevados luego de que la influencia de los EF ha sido evaluada. Para cumplir estos objetivos un número de técnicas han sido descritas (Pfeiffer et al., 2008). Desde aproximaciones frecuentistas, donde un resumen global de la fuerza de los efectos para cada región es provista en adición a los coeficientes de regresión para cada uno de los EF en el modelo (ej: Walter, Martin Taylor y Marrett, 1999), hasta modelos mixtos Bayesianos completos donde efectos aleatorios correlacionados son determinados a nivel de región (ej: Toledano, Jarup, Best, Wakefield y Elliott, 2001 y Jarup, Best, Toledano, Wakefield y Elliott, 2002). La aproximación Bayesiana es la que interesa y se trata aquí.

Un número considerable de literatura se ha desarrollado alrededor de la aproximación Bayesiana para modelar conteos de enfermedades dentro de regiones definidas arbitrariamente (Besag et al., 1991; Marshall, 1991; Clayton y Benardinelli, 1992; Breslow y Day, 1993; Lawson, 1994; Ghosh, Natarajan, Stroud y Carlin, 1998; Lawson et al., 2001, Lawson 2009 y 2013).

#### **MODELO (BYM) BESAG, YORK Y MOLLÍE.**

En este modelo para el riesgo relativo, efectos aleatorios área específicos se descomponen dentro de un componente que toma en cuenta los efectos que varían de una manera estructurada en el espacio (heterogeneidad correlacionada) y un componente

que modela los efectos que varían de una manera desestructurada entre áreas (heterogeneidad no correlacionada).

El modelo introducido por Clayton y Kaldor (1987) y desarrollado por Besag et. al. (1991), es formulado como sigue:

$$y_i \sim \text{Poisson}(e_i \theta_i),$$

$$\log \theta_i = \alpha + v_i + u_i$$

Donde  $\alpha$  es el nivel general de riesgo relativo,  $u_i$  es la heterogeneidad correlacionada y  $v_i$  es la heterogeneidad no correlacionada.

Los modelos Bayesianos requieren especificación de la distribución a priori para los efectos aleatorios. El modelo de distribución para la heterogeneidad no correlacionada es

$$v_i \sim N(0, \tau_v^2)$$

Para el componente de heterogeneidad estructurada, se usa una estructura espacial correlacionada, donde la estimación del riesgo en cada área depende de áreas vecinas. Se usa el modelo CAR propuesto por Besag et al. (1991)

$$[u_i | u_j, i \neq j, \tau_u^2] \sim N(\bar{u}_i, \tau_i^2)$$

Donde

$$\bar{u}_i = \frac{1}{\sum_j \omega_{ij}} \sum_j u_j \omega_{ij},$$

$$\tau_i^2 = \frac{\tau_u^2}{\sum_j \omega_{ij}},$$

$\omega_{ij} = 1$  si  $i, j$  son adyacentes (o 0 si no lo son)

Los parámetros  $\tau_v^2$  y  $\tau_u^2$  controlan la variabilidad de  $v$  y  $u$ . En el análisis Bayesiano completo, distribución a priori pueden ser especificadas para estos parámetros. Una opción común es considerar una distribución gamma para ambos, como sugiere Bernardinelli et al. (1995).

#### 4.8. ANÁLISIS ECOLÓGICO

Estudios de regresión ecológicos tienen una larga historia en epidemiología (Richardson, Jackson & Best, 2005), el objetivo es examinar la variación geográfica de los grupos de población expuestos a variables ambientales (que pueden ser medidas en el aire, agua o suelo), socioeconómicas y demográficas (tales como la raza y los ingresos), o factores de estilo de vida (como el tabaquismo y la dieta) en relación con los resultados de salud medidos en una escala geográfica (ecológica), (Elliott y Wartenberg, 2004).

El análisis ecológico está relacionado a la construcción de mapas de enfermedades. El objetivo del estudio ecológico es describir la relación entre la variación geográfica de la incidencia o mortalidad y covariables explicativas medidas en grupos más que individualmente (Lawson, 2003).

Estudios ecológicos suelen utilizar datos de observación recogidos en grupos definidos por áreas geográficas como países, regiones, departamentos o radios censales, que derivan de forma rutinaria a través de diversos tipos de censo, registros nacionales o los sistemas de salud públicos (Richardson, Jackson y Best, 2005; Lawson, 2009).

Los estudios ecológicos son usados en conjunto con estudios descriptivos simples de la variación geográfica en un intento de determinar cuánto de la variación en la tasa de enfermedad está asociado con la variación en la exposición (Lawson, 2003).

Un estudio ecológico puede ser hecho en base a datos georreferenciados, pero el foco no es por lo general la distribución espacial, sino la relación entre la variable dependiente y las covariables medidas (Lawson, 2003).

El análisis espacial de los datos epidemiológicos que se deben realizar en un estudio ecológico se puede subdividir en varios pasos: i) Identificación de un posible problema de salud, ii) Verificación de que un patrón espacial observado de casos de la enfermedad difiere considerablemente del patrón esperado y iii) Análisis ecológico: basado en el conocimiento de la distribución de la enfermedad y teniendo en cuenta la distribución espacial de los factores de riesgo, se formulan ciertas hipótesis sobre la causa de la distribución de la enfermedad y se evalúan por medio de métodos estadísticos (Kistemann et al., 2002).

Un problema importante que afecta a los análisis ecológicos es el sesgo ecológico. Este sesgo surge cuando la asociación observada en el grupo o nivel ecológico se aplica a la asociación entre las variables correspondientes a nivel individual (Elliott y Wartenberg, 2004; Lawson, 2009).

En términos generales, cuanto menor sea el área en la que se recolectan los datos de enfermedades y ambientales, mayor es la posibilidad de hacer inferencias precisas,

porque las medias en grandes áreas pueden introducir sesgo ecológico considerable (Rinaldi et al., 2006).

A pesar de ello, los estudios ecológicos tienen algunas ventajas sobre otros estudios, ya que incluyen la posibilidad de estudiar una población grande a un bajo costo y de abordar cuestiones de salud ambiental que podrían ser difíciles de abordar a nivel individual; son particularmente útiles cuando las mediciones individuales de la exposición son difíciles o imposibles de obtener (por ejemplo, la contaminación del aire) o se miden de manera imprecisa (por ejemplo, la ingesta diaria y la exposición al sol), (Lawson, 2003).

## MODELO ESTADÍSTICO

La aproximación simple al análisis ecológico es el uso de modelos de regresión para tasas de enfermedad que solo permiten variación de Poisson.

Si  $\{y_1, \dots, y_m\}$  es el número de eventos observados y  $\{e_1, \dots, e_m\}$  el número esperado de eventos para cierta enfermedad en  $m$  áreas de la región de interés. Asumimos que  $y_i$  se distribuye de acuerdo a una distribución de Poisson con media  $u_i = e_i \theta_i$ , donde  $\theta_i$  es el riesgo relativo en el área  $i$ . Como mencionamos previamente, la estimación de la verosimilitud máxima saturada de  $\theta_i$  esta dada por la SMR. El modelo puede extenderse para incluir un set de variables explicativas  $x_1, x_2, \dots, x_p$  en una formulación log - lineal.

$$\log u_i = \log e_i + \sum_{j=1}^p \beta_j x_{ij}$$

Este modelo puede incluir muchos factores ecológicos (covariables  $x_j$ ), pero, en muchas situaciones, la variación no explicada por las variables ecológicas podría exceder la que se espera a partir del modelo de Poisson y conduce a sobredispersión.

Variabilidad extra-Poisson espacialmente estructurada puede ser fácilmente modelada asumiendo que la verdadera distribución del riesgo relativo toma una forma log - normal con componentes aditivos.

Por lo tanto, un parámetro para sobredispersión se añade al modelo. El parámetro de sobredispersión se refiere al grado de heterogeneidad entre los verdaderos riesgos relativos, dadas las variables en el modelo.

Fuentes de variabilidad extra-Poisson no estructuradas y estructuradas pueden explicarse mediante la adopción de un modelo donde la variación extra-Poisson se descompone en dos componentes.

El primer componente es la variación espacialmente no estructurada extra-Poisson (heterogeneidad no estructurada) y el segundo componente varía suavemente a través de las áreas (heterogeneidad estructurada). La introducción de estos términos en el modelo representa una forma de controlar las covariables no medidas. El modelo podría escribirse de la siguiente manera:

$$\log(\theta_i) = \alpha + x_i'\beta + u_i + v_i$$

Donde  $\alpha$  es una constante,  $x_i$  representa un vector de variables explicativas o covariables en la región  $i$ ,  $\beta$  es un vector de parámetros, y los componentes de la variación, heterogeneidad no estructurada y espacialmente estructurada, son representados por  $v_i$  y  $u_i$  respectivamente. Este modelo fue introducido por Clayton y Kaldor (1987) y desarrollado por Besag et al. (1991) y Clayton et al. (1993). En este trabajo, Clayton et al. (1993) recalca la importancia de incluir un término espacialmente correlacionado ( $U_i$ ) en análisis ecológicos para permitir la variación no observada.

Ejemplos de estudios de regresión ecológica que han implementado modelos de esta forma incluyen Clayton y Bernardinelli (1992), que analizan los efectos de la urbanización sobre la mortalidad por cáncer de mama en Cerdeña, y Richardson, Monfort, Green, Draper y Muirhead (1995) que analizan la variación geográfica de la leucemia infantil en el Reino Unido en relación con la radiación natural. Ver también Bernardinelli et al. (2000) y Bernardinelli et al. (1999) para continuar el debate. Para una amplia gama de modelos estadísticos que se proponen en el contexto de la regresión ecológica ver Biggeri et al. (1999). Ellos discutieron los modelos mencionados anteriormente, modelos jerárquicos de efectos aleatorios y modelos no paramétricos para este tipo de análisis.

Modelos jerárquicos bayesianos espaciales con diferentes distribuciones para modelar la variable respuesta (Poisson, Poisson inflado con ceros, binomial y binomial negativo) fueron utilizados en los siguientes trabajos con el fin de realizar mapas de SMR y/o el análisis ecológico, Durr et al. (2005) tienen como objetivo identificar áreas de alta prevalencia y explicar esto en términos de covariables ambientales utilizando un conjunto de datos históricos de una encuesta de matadero para fasciolosis realizado en Victoria, Australia, Yu et al. (2008) relacionan la destrucción en los puntos de venta de alcohol y los cambios en la tasa de violencia en el vecindario, Willgert et al. (2011) estudian el efecto de la vacunación sobre la incidencia espacial de lengua azul (BTV-8) y el riesgo relativo de la enfermedad asociada en Suiza, Sartorius et al. (2011) identifican y mapean grupos de alto riesgo de mortalidad infantil, así como examinan el

impacto de una serie de factores determinantes de la mortalidad infantil, Puigpinós-Riera (2011) analizan los patrones de mortalidad por cáncer y su relación con la privación socioeconómica en áreas pequeñas en 11 ciudades españolas, Hoffmann et al. (2014) hicieron frente a los problemas asociados con razones de mortalidad estandarizadas para áreas pequeñas suavizando mediante el modelo bayesiano propuesto por Besag, York y Mollié e utilizaron un análisis de regresión ecológica para evaluar la asociación entre la privación social y la mortalidad.

Detalles adicionales e ilustraciones pueden encontrarse en varios textos y libros como, Mollié (1996), Best et al. (1999), Wakefield et al. (2000), Lawson (2001), Banerjee et al. (2004), Waller and Gotway (2004), Waller (2005), Carlin and Louis (2008) y Lawson (2009, 2013).

#### **4.9. Residuos y bondad de ajuste**

El logro de la convergencia de los algoritmos MCMC no produce necesariamente buenos modelos. Si está mal especificado un modelo, entonces será de uso limitado. Hay muchas cuestiones relacionadas con la bondad de ajuste del modelo que deben ser motivo de preocupación en la evaluación de los modelos de datos geo-referenciados de enfermedades. Se tratan cuestiones generales relacionadas con el uso de medidas de bondad del ajuste (MBA) y diagnósticos residuales.

##### **4.9.1. MEDIDAS DE BONDAD DE AJUSTE PARA MODELOS**

El criterio de bondad de ajuste varía dependiendo de las propiedades del criterio y la naturaleza del modelo. En modelos lineales generalizados (GLM) convencionales con efectos fijos, la desviación es una medida importante (McCullagh y Nelder, 1989). Usualmente esta medida de adecuación del modelo compara un modelo fijado y uno saturado. Esto basado en la diferencia entre el log de la verosimilitud de los datos entre los modelos:

$$D = -2[l(y|\hat{\theta}_{fit}) - l(y|\hat{\theta}_{sat})]$$

El modelo saturado tiene un solo parámetro por observación. A menudo se utiliza una medida relativa de ajuste para comparar la desviación y se utiliza el cambio en la desviación entre el modelo 1 y el modelo 2:

$$\Delta D = -2[l(y|\hat{\theta}_1) - l(y|\hat{\theta}_2)]$$

Por lo tanto, la verosimilitud saturada se cancela en esta comparación relativa. La desviación se utiliza en la medida de bondad de ajuste en el modelado bayesiano, pero por lo general sin referencia a un modelo saturado.

Una desventaja del uso de la desviación directamente es que no mide el grado de parametrización en el modelo: un modelo puede aproximarse más a los datos por incremento en el número de parámetros. Por lo tanto, se han hecho intentos para penalizar la complejidad del modelo. Un ejemplo de esto es el criterio de información de Akaike (AIC). Este se define como

$$AIC = -2[l(y|\hat{\theta}_{fit}) + 2p]$$

donde  $p$  es el número de parámetros.

El segundo término actúa como sanción para la sobre parametrización del modelo. La idea es que a medida que más parámetros se añaden, entonces de mejor manera el modelo se aproximará a los datos.

Para equilibrar esto, se asume la pena ( $2p$ ). Por lo tanto, el ajuste es penalizado con una función lineal del número de parámetros. Modelos parsimoniosos deben resultar del uso de tal penalización. Esto es ampliamente utilizado para los modelos de EF y es la base del criterio de información de desviación que se explica a continuación.

Otra variante que se utiliza comúnmente como criterio para la elección del modelo es el criterio de información bayesiano (BIC). Este es ampliamente utilizado en los modelos bayesianos y jerárquicas. Se aproxima asintóticamente un factor de Bayes. Se define como

$$-2l(y|\hat{\theta}_{fit}) + p \ln m.$$

En un modelo con log-verosimilitud  $l(\theta)$  el valor de AIC o BIC puede estimarse a partir de la salida de un algoritmo MCMC por

$$AIC = -2\hat{l}(\theta) + 2p$$

$$BIC = -2\hat{l}(\theta) + p \ln m,$$

donde  $p$  es el número de parámetros,  $m$  es el número de datos de punto y

$$\hat{l}(\theta) = \frac{1}{G} \sum_{g=1}^G l(y|\theta^g),$$

el promedio del log-verosimilitud sobre  $G$  muestras posteriores de  $\theta$ .

Alternativamente una estimación posterior de  $\hat{\theta}$  (tal como la expectativa posterior) puede ser primero calculada y entonces sustituida dentro de AIC o BIC. Leonard y Hsu (1999) proporcionan comparaciones de estas medidas en una variedad de ejemplos. Una de las desventajas de la AIC o BIC es que, en los modelos con efectos aleatorios, es difícil decidir cuántos parámetros se incluyen dentro del modelo. Por ejemplo, un efecto de nivel de unidad se especifica como  $v_i \sim N(0, \tau)$ . En este caso hay un parámetro varianza, pero también hay un valor separado de  $v$  para cada elemento. Por lo tanto, tenemos potencialmente super-saturación del espacio de parámetros ( $p > m$ ) si contamos  $m$   $v_i$ s así como  $\tau$ . Debería ser la parametrización 1 o  $m + 1$  o en algún lugar entre estos valores? Este dilema no se plantea con modelos de coeficientes aleatorios donde, por ejemplo, en un contexto de regresión en el que se puede tener un vector de longitud  $p$  de parámetros,  $\beta$  se dice, que puede tener en el caso, simple  $p$  varianzas.

#### CRITERIO DE INFORMACIÓN DE DESVIACIÓN

El “criterio de información de desviación” (DIC, por sus siglas en inglés) ha sido propuesto por Spiegelhalter, Best, Carlin y Van Der Linde (2002) y es extensamente utilizado en modelado Bayesiano. Este se define como

$$DIC = 2E_{\theta|y}(D) - D[E_{\theta|y}(\theta)],$$

Donde  $D(\cdot)$  es la desviación del modelo e “ $y$ ” es el dato observado. Note que el DIC es basado en la comparación del promedio de la desviación ( $\bar{D} = -2 \sum_{g=1}^G l(y|\theta^g)/G$ ) y la desviación de las estimaciones de los parámetros posteriores esperados,  $\hat{\theta}$  se dice: ( $\hat{D}(\hat{\theta}) = -2l(y|\hat{\theta})$ ). Para cualquier valor de parámetro de muestra  $\theta^g$  la

desviación es justo  $\widehat{D}(\theta^g) = -2l(y|\theta^g)$ . El número efectivo de parámetros ( $pD$ ) se estima como  $\widehat{pD} = \overline{D}(\widehat{\theta})$  y entonces  $DIC = \overline{D} + \widehat{pD} = 2\overline{D} - \widehat{D}(\widehat{\theta})$ . Desafortunadamente en algunas situaciones  $\widehat{pD}$  puede ser negativo (ya que puede suceder que  $\widehat{D}(\widehat{\theta}) > \overline{D}$ ). Inestabilidad en  $pD$  puede conducir a problemas en el uso del DIC. Por ejemplo, modelos mixtos o más simples y modelos con modos múltiples pueden “engañar” la estimación del  $pD$  a causa de sobredispersión en tales modelos (cuando los componentes no se estiman correctamente) conduciendo a  $D(\theta) > D$ . Sin embargo, también es cierto que la elección inapropiada de hiper-parámetros para las variaciones de los parámetros en los modelos jerárquicos puede conducir también a la inflación, como transformaciones no lineales (tales como cambiando de un modelo de Gauss a un modelo log - normal). En tales casos, a veces es más seguro calcular el número efectivo de parámetros desde la varianza posterior de la desviación. Gelman et al. (2014, p. 172) propone el estimador  $\widehat{pD} = \frac{1}{2} \frac{1}{G-1} \sum_{g=1}^G (\widehat{D}(\theta^g) - \overline{D})^2$ . Este valor puede también ser calculado a partir de la muestra del resultado de una cadena. (También está disponible directamente en R2WinBUGS.) Un estimador alternativo de la varianza es directamente disponible desde el resultado  $\widehat{var}(D) = \frac{1}{G-1} \sum_{g=1}^G (\widehat{D}(\theta^g) - \overline{D})^2 = 2\widehat{pD}$ . De ahí que un DIC con base en esta última estimación de la varianza es justo  $DIC = \overline{D} + \widehat{var}(D)$ . Note que la desviación predictiva esperada (EPD:  $D_{pr}$ ) es una medida alternativa de adecuación del modelo y esta es basada en la capacidad predictiva del modelo ajustado fuera de la muestra. La cantidad también puede estimarse aproximadamente como  $\widehat{D}_{pr} = 2\overline{D} - \widehat{D}(\widehat{\theta})$ .

## PÉRDIDA PREDICTIVA POSTERIOR

Gelfand y Ghosh (1998) propusieron un enfoque basado en la función de pérdida que emplea la distribución predictiva para modelar la adecuación. El enfoque esencialmente compara los datos observados de los datos predichos desde el modelo ajustado. Se define el dato  $i$  predictivo como  $y_i^{pr}$ . Tenga en cuenta que los datos de predicción pueden ser fácilmente obtenidos a partir de una muestra posterior convergente. Dados los parámetros actuales en iteración  $j$  :  $\theta^{(j)}$  se dice, entonces

$$p(y_i^{pr} | y) = \int p(y_i^{pr} | \theta^{(j)}) p(\theta^{(j)} | y) d\theta^{(j)}$$

De aquí que la iteración  $j$  th puede producir  $y_{ij}^{pr}$  desde  $p(y_{ij}^{pr} | \theta^{(j)})$ . El valor predictivo resultante tiene distribución marginal  $p(y_{ij}^{pr} | y)$ . Para una distribución de Poisson, esto simplemente requiere generación de conteos como  $y_{ij}^{pr} \leftarrow Pois(e_i \theta_i^{(j)})$ .

Una función de pérdida se asume donde  $L_0(y, y^{pr}) = f(y, y^{pr})$ . Una opción conveniente de la pérdida podría ser la pérdida de error al cuadrado por el que se define la pérdida como:  $L_0(y, y^{pr}) = (y - y^{pr})^2$ .

Funciones de pérdida alternativas podrían proponerse como la pérdida de error absoluto o formas más complejas (cuantiles). Una medida bruta global de pérdida a través de los datos se produjo por la pérdida media en todos los ítems: el error cuadrático medio de predicción (MSPE, por sus siglas en inglés) es simplemente un promedio del ítem error de pérdida al cuadrado:

$$MSPE_j = \sum_i (y_i - y_{ij}^{pr})^2 / m$$

y

$$MSPE = \sum_i \sum_j (y_i - y_{ij}^{pr})^2 / (G \times m),$$

donde  $m$  es el número de observaciones y  $G$  es el tamaño de la muestra muestreado. Una alternativa podría ser especificar un error absoluto:

$$MAPE_j = \sum_i |y_i - y_{ij}^{pr}| / m$$

y

$$MAPE = \sum_i \sum_j |y_i - y_{ij}^{pr}| / (G \times m),$$

Gelfand y Ghosh (1998) proponen una forma más sofisticada

$$D_k = \frac{k}{k+1} A + B$$

$$= \frac{k}{k+1} \sum_{i=1}^m (y_i - y_i^{\bar{pr}})^2 + \frac{1}{m_p} \sum_{i=1}^m \sum_{j=1}^{m_p} (y_{ij}^{pr} - y_i^{\bar{pr}})^2$$

$$\text{donde } y_i^{\bar{pr}} = \sum_{j=1}^{m_p} y_{ij}^{pr} / m_p$$

y  $m_p$  es el tamaño de la muestra de predicción (por lo general  $G = m_p$ ). Aquí  $k$  puede ser elegido para ponderar los diferentes componentes. Para  $k = \infty$ , entonces  $D_k = A + B$ . La elección de  $k$  no suele afectar al orden de ajuste del modelo. Cada componente mide una característica diferente del ajuste: A representa la falta de ajuste y B grado de suavidad. El modelo con más bajo  $D_k$  (o MSPE y/o MAPE) será preferible (Lawson, 2009).

Tenga en cuenta que la predicción de los datos se puede obtener fácilmente de fórmulas modelo en WinBUGS: en el caso de Poisson con datos  $y[i]$  observados y datos predictivos  $ypred[i]$ , tenemos para el  $i$ -ésimo elemento:

$$y[i] \sim \text{dpois}(\mu[i])$$

$$ypred[i] \sim \text{dpois}(\mu[i])$$

Como faltan los valores predichos, estos tendrían que ser inicializados.

Además, es posible considerar medidas basadas en predicciones como el diagnóstico de convergencia. La medida  $\widehat{D}(\theta^{(j)})$  o  $MSPE_j$  puede ser monitoreada usando diagnósticos de simple o múltiples cadenas. Tenga en cuenta que para cualquier modelo para el que una contribución de unidad de verosimilitud está disponible, es posible calcular una medida basada en la varianza como  $\widehat{D}(\theta^{(j)})$ . Por lo tanto para el proceso de punto (evento caso), así como verosimilitudes basadas en conteos, las medidas de desviación están disponibles mientras que una medida basada en residuos (como MSPE) es más difícil de definir para un evento con dominio espacial.

#### 4.9.2. RESIDUOS GENERALES

El análisis de los residuos y las funciones de resumen de los residuos constituye una parte fundamental de la evaluación de la bondad de ajuste del modelo en cualquier área de aplicación estadística.

En el caso de la cartografía de enfermedades no es la excepción, aunque el análisis residual total rara vez se presenta en trabajos publicados en el área.

A menudo, las medidas de bondad de ajuste son funciones agregadas de trozos de residuos, mientras que las medidas relativas a los residuos individuales también están disponibles.

Una variedad están disponibles cuando el análisis residual total se va a realizar.

Definimos un trozo de residuo como la diferencia estandarizada entre el valor observado y el valor del modelo ajustado.

Por lo general, la normalización se basa en una medida de la variabilidad de la diferencia entre los dos valores.

Es práctica común especificar el residuo como

$$r_i = y_i - \hat{y}_i$$

O

$$r_i^s = r_i / \sqrt{\widehat{var}(r_i)}$$

Donde  $\hat{y}_i$  es un valor fijado bajo un modelo dado. Cuando se consideran modelos espaciales complejos, a menudo es más fácil examinar los residuos, tales como  $\{r_i\}$  usando métodos Monte Carlo. De hecho, es sencillo de implementar un enfoque *bootstrap* paramétrico (PB) para el diagnóstico residual para los modelos de verosimilitud (Davison y Hinkley, 1997). El caso más simple en conteos por área, es cuando por cada área una cuenta observada puede ser comparada con un recuento ajustado. En general cuando se asumen modelos de verosimilitud de Poisson con  $y_i \sim Pois\{e_i \theta_i\}$  entonces es fácil de emplear un PB mediante la generación de un conjunto de recuentos simulados  $\{y_{ij}^*\} j = 1, \dots, J$ , desde una distribución de Poisson con media  $e_i \hat{\theta}_i$ . De esta manera, una clasificación del área, y por lo tanto el valor de p, se pueden calcular mediante la evaluación del grado de los residuos dentro del conjunto de  $J + 1$  residuales:

$$\{y_i - e_i \hat{\theta}_i; \{y_{ij}^* - e_i \hat{\theta}_i\}, j = 1, \dots, J\}.$$

Denotar los residuos estandarizados observados como  $r_i^s$ , y el simulado como  $r_{ij}^s$ . Tenga en cuenta que ahora es posible comparar las funciones de los residuos, así como hacer

comparaciones directas. Por ejemplo, en un contexto espacial, puede ser apropiado examinar la autocorrelación espacial de los residuos observados. Esto puede proporcionar evidencia de la falta de ajuste del modelo. Por lo tanto, una evaluación Monte Carlo del grado de autocorrelación residual puede ser hecha por comparación de la estadística de Moran's I para los residuos observados, dicese,  $M(\{r_i^s\})$ , a la encontrada para los residuos de conteo simulados  $M(\{r_{ij}^s\})$ , donde  $M(\{u\}) = \frac{u^T W u}{u^T u}$  donde  $u_i = r_i / \sqrt{\text{var}(r_i)}$  y  $r_i = (y_i - e_i \hat{\theta}_i)$  y W es una matriz de adyacencia. Note que  $E[M(\{u\})]$  puede no ser cero, por lo que sería importante tener en cuenta este hecho en cualquier evaluación de autocorrelación residual.

## RESIDUOS BAYESIANOS

En un entorno bayesiano es natural considerar la versión adecuada de  $r_i = y_i - \hat{y}_i$ . Carlin y Louis (2000) describen un residuo bayesiano

$$r_i = y_i - \frac{1}{G} \sum_{g=1}^G E(y_i | \theta_i^{(g)})$$

Donde  $E(y_i | \theta_i)$  es el valor esperado de la distribución predictiva a posteriori, y (en el contexto de muestreo MCMC)  $\{\theta_i^{(g)}\}$  es un set de parámetros muestreados desde la distribución a posteriori.

En el caso de modelos de conteos por áreas, con verosimilitud de Poisson y esperados  $e_i \theta_i$ , este residuo puede ser aproximado, donde una tasa constante de área se asume:

$$r_i = y_i - \frac{1}{G} \sum_{g=1}^G e_i \theta_i^{(g)}.$$

Esto da un promedio residual sobre la muestra posterior. Una probabilidad computacional alternativa es promediar la muestra  $\{\theta_i^{(g)}\}$ ,  $\hat{\theta}_i = \frac{1}{G} \sum_{g=1}^G \theta_i^{(g)}$ , es decir, para producir un valor posterior esperado de  $y_i$ , es decir  $\hat{y}_i = e_i \hat{\theta}_i$ , y para formar  $r_i = y_i - \hat{y}_i$ . Una futura posibilidad es formar simplemente  $r_i$  en cada iteración de un muestreador a posteriori y promediar estos sobre la muestra convergente (Spiegelhalter et al., 1996). Estos residuos pueden proporcionar medidas puntuales de bondad de

ajuste, así como MBA globales, (como el error cuadrático medio (MSE):  $\frac{1}{m} \sum_{i=1}^m r_i^2$ ) y puede evaluarse utilizando métodos de Monte Carlo. Para fines de exploración podría ser útil estandarizar los residuos antes del examen, aunque esto no es esencial para la evaluación de Monte Carlo. Para proporcionar una evaluación de Monte Carlo de la unidad de comportamiento individual residual sería necesaria una simulación de Monte Carlo repetida de muestras independientes de la distribución predictiva. Esto se puede lograr tomando J muestras de MCMC convergente con brechas de longitud p, donde p es lo suficientemente grande como para asegurar la independencia. Clasificar los residuos en un conjunto combinado (J + 1), puede utilizarse para proporcionar un valor de p Monte Carlo para cada unidad.

### **INTERPRETACIÓN DE LOS RESIDUOS EN UN AJUSTE BAYESIANO**

Diagnósticos basados en residuos serán indicativos de una variedad de características del modelo. ¿Qué se debe esperar de los residuos de un modelo adecuado? En general, cuando un modelo se ajusta bien, sería de esperar que los residuos del modelo tengan un número de características. En primer lugar por lo general deben ser simétricos y centrados en torno a cero. Es evidente que la normalización de la varianza debe producir una aproximación más cercana a este, pero en general esto sólo puede ser aproximado. En segundo lugar no deben mostrar ninguna estructura particular y deben aparecer razonablemente al azar. Sin embargo, en cuanto a la distribución, no está claro que los residuos deben tener una forma gaussiana de media cero (como se sugiere por el uso de parcelas de cuantiles normales en Carlin y Louis, (2000)).

---

# *METODOLOGÍA Y ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS.*

## **CAPÍTULO 5:**

### **5. METODOLOGÍA Y ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS.**

#### **5.1. Región de estudio**

##### Metodología

La región de estudio fue definida tomando los radios censales de la Provincia de Río Negro (Figura 3, apéndice A), que coinciden en su mayoría con el área programa Hidatidosis, menos los que solapan con el río Negro y los del norte de éste. También se eliminaron las partes de los radios censales que cubren los cuerpos de agua como el lago Nahuel Huapi y el embalse el Chocón.

#### **5.2. Definición y construcción de variables**

##### Metodología

Para agrupar todas las variables que se describen a continuación, se utilizó un archivo shape (<http://resources.arcgis.com/es/help/main/10.1/index.html#//005600000002000000>) de radios censales 2010, provisto por la Infraestructura de Datos Espaciales de la provincia de Río Negro (<http://www.ide.rionegro.gov.ar/index.php?catID=186>).

#### **CASOS**

Los casos incluidos en el estudio se encuentran en la base de datos provista por el departamento de zoonosis dependiente del Ministerio de Salud de Río Negro y representan una serie temporal (1996 a 2010) de las coordenadas geográficas Lat/Lon, Datum WGS-84, del sitio probable de exposición de chicos de entre 0 a 14 años con diagnóstico de EQ confirmada, siendo diagnosticados mediante encuestas con

ultrasonografía en población asintomática (escolares de 7 a 14 años) o por haber presentado sintomatología clínica.

La definición de caso se basa en la establecida en las normas nacionales para el control de la hidatidosis, siendo

Caso sospechoso:

- a) Persona sintomática con presencia de masa quística localizada en distintos órganos y sistemas, con más frecuencia en hígado y pulmón, y asociado con aspectos epidemiológicos de la enfermedad (lugar de origen, contacto con perros, existencia de otros familiares con hidatidosis)
- b) Persona positiva a tamizaje ecográfico o serológico, y asociada con aspectos epidemiológicos de la enfermedad (lugar de origen, contacto con perros, existencia de otros familiares con hidatidosis)

Se considera caso confirmado al caso sospechoso con diagnóstico por imágenes (radiografía, ecografía y/o tomografía axial computarizada) y/o pruebas serológicas (ELISA/Westernblot).

En Argentina, asimismo, la confirmación parasitológica requiere la visualización directa por microscopía de protoescolices o ganchos del parásito, restos de membranas y estudio histopatológico de la pieza extraída por cirugía

Los casos fueron agregados por radio censal (Figura 4, apéndice A), mediante intersección de sus coordenadas con los radios censales.

## **POBLACIÓN**

La población fue calculada en forma aproximada para los radios censales a partir de los censos 2001 y 2010. Representa solo la población rural, habiéndose extraído la población de los radios censales considerados urbanos con el software QGIS 2.8.1. Una aproximación lineal se realizó para calcular los años 2002 a 2009.

## **ESPERADOS**

Los esperados por radio censal fueron calculados como se mencionó previamente en el texto (ver sección 4.1.2.) mediante el software R (<https://www.r-project.org/>).

### **5.2.1. COVARIABLES**

## Metodología

Las covariables fueron construidas a partir de productos provenientes de Sensores Remotos y Sistemas de Información Geográfica.

### **TEMPERATURA**

Se obtuvieron productos del sensor MODIS (MOD11A2 diurnos) a bordo del Satélite Terra. Los productos MOD11 (MODIS Land Surface Temperature and Emissivity, s.f.) brindan estimaciones de la temperatura y emisividad diurna y nocturna de las coberturas terrestres (LST), la temperatura se mide en grados Kelvin al contacto entre la cobertura y la atmósfera. Estos productos son producidos con una resolución espacial de 1000 m con base diaria y cada 8 días (Mas, 2011).

El producto MOD11A2 fue descargado vía MRTWeb 2.0 (MRTWeb, s.f.) que combina funciones de búsqueda, visualización y selección de USGS Global Visualization Viewer (Glovis) y funciones para armar mosaicos, selección de subconjunto espacial, selección de bandas, re - proyección, re - muestreo y selección de formato de *MODIS Reprojection Tool* (MRT). La interfaz MRTWeb organiza funcionalidades Glovis y MRT en tres pestañas principales: Selección, Proceso, y Descarga. Los trabajos enviados se ejecutan con el software de procesamiento MRT 4.0 a través de múltiples servidores DAAC LP. El conjunto de datos de salida queda disponible en un directorio de trabajo específico ftp para descargar por el usuario.

Se descargaron 437 imágenes MOD11A2 para la región de estudio, desde el día juliano 001 de 2001 cada 8 días hasta el día juliano 361 de 2010, en proyección geográfica Lat/Lon, Datum WGS-84.

Las imágenes fueron procesadas con el software ENVI 4.8, donde primero se agruparon en un *Layer Stacking* y en IDL se realizó la interpolación temporal para corregir los valores faltantes (0 en nuestro caso) por el de la fecha siguiente y se pasaron a grados Celsius. Los valores que durante las 437 imágenes tuvieron datos faltantes se enmascararon con *No data* (NAN), a fin de no afectar la estadística zonal que se calculó posteriormente. Con el *Layer Stacking* con las imágenes interpoladas y enmascaradas se procede a realizar estadísticos a través de la herramienta *Sum Data Band*. Con esta herramienta se obtiene una imagen (dos bandas) con la media y desvío estándar para cada píxel, calculadas desde los píxeles de la serie temporal (437 imágenes).

Luego, dentro del software QGIS 2.8.1 se obtiene mediante la herramienta Estadísticas de zona, la media y desvío estándar para cada radio censal.

## TEMPERATURA POR RANGOS

Se clasificó la temperatura de los radios censales en 5 clases (Figura 5, apéndice A) y de acuerdo a su correlación con la SMR de hidatidosis en chicos de 0 a 14 años, se seleccionó el rango que va de 17.2 a 19.9 °C (el de mayor correlación, ver figura 13 y 14, apéndice A) para crear una nueva variable, donde se asignó, valor = 1, para los radios censales que están en el rango descripto y valor = 0, para el resto (Figura 6, apéndice A).

## VEGETACIÓN

El NDVI, Figura 7, apéndice A, fue también obtenido desde el sensor MODIS a bordo del satélite Terra. Mediante la herramienta MRTWeb, se obtuvieron 110 imágenes del producto MOD13A3 (MOD13A3, s.f.), índice de vegetación mensual, 1 km de resolución espacial, para los años 2001 a 2010, en proyección geográfica Lat/Lon, Datum WGS-84.

Las imágenes se agruparon en un *Layer Stacking*, se computó la media y el desvío estándar de la misma forma que para la temperatura y al final se aplicó una máscara con NAN para la superficie del mar. Luego se calcularon los estadísticos de zona en QGIS.

## PRECIPITACIÓN

Un promedio temporal de precipitación entre 2000 - 2010, basado en el producto 3B43 v7, del Tropical Rainfall Measuring Mission Project (TRMM) y otras fuentes mensuales de precipitación, fue obtenido desde la aplicación Web Giovanni (Giovanni The Bridge Between Data and Science, s.f.). La imagen resultante tiene una resolución espacial de 0.25 grados y resolución temporal mensual. La proyección se definió de igual forma que las variables anteriores. Los promedios zonales fueron calculados como se mencionó previamente para la temperatura y vegetación.

## ALTURA

La altura, fue obtenida a través de un modelo digital de elevación de 90 metros de la Shuttle Radar Topographic Mission (SRTM) de la NASA. El mosaico para la región de estudio fue provista por la Comisión Nacional de Actividades Espaciales (CONAE), en proyección geográfica Lat/Lon, Datum WGS-84, resolución espacial 0.000833 grados.

La estadística zonal fue calculada luego en QGIS.

## REGIONES NATURALES

De acuerdo a Godagnone y Bran (2009) que tratan las características generales de la provincia de Río Negro en cuanto se refiere a los aspectos climáticos, geológicos, hídricos, geomorfológicos, edáficos y de vegetación, agrupándolas en tres Regiones Naturales (Patagonia Andina, Patagonia Extra Andina Occidental y Patagonia Extra Andina Oriental) y de acuerdo con Villagra y Giraud (2013) que basados en diversos trabajos realizados sobre características productivas, biofísicas y socioeconómicas en la provincia, diferencian dos regiones en las que se desarrolla la producción ovina (Patagonia Extra Andina Occidental y Monte Austral Rionegrino), se decide contrastar la región denominada Patagonia Extra Andina Occidental (PEAO) con el resto del territorio, teniendo en cuenta que esta región que se menciona por ambos autores, tiene características diferenciales (que podrían influir en el ciclo biológico del *Echinococcus granulosus*) y además, a la exploración visual se solapa con la mayoría de casos de Hidatidosis.

Por lo expuesto se construye una variable binaria, donde cada radio censal toma un valor de 1, si su superficie intersecta en un 70% o más con la de la región PEAO y un valor de 0 de otro modo (Figura 8, apéndice A).

Las características más sobresalientes de la Región PEAO son, según Godagnone y Bran (2009), en primer término, el clima que varía de subhúmedo en el límite Oeste a semiárido y árido hacia el Este; las geoformas más abundantes son bajadas, serranías rocosas, mesetas lávicas y depresiones sin salida. Los suelos pertenecen, mayoritariamente, a los Subórdenes de los Argides y Ortentes, en menor cantidad se encuentran Xeroles y Ustoles. La vegetación está caracterizada por el predominio de estepas arbustivo-graminosas bajas.

### 5.2.2. SELECCIÓN DE COVARIABLES

#### Metodología:

Se evaluó la correlación entre covariables, de las que se correlacionaron en más de un 90% se seleccionaron aquellas de mayor valor biológico – epidemiológico, que por sus características pudiesen influir en el ciclo biológico – epidemiológico de la enfermedad.

Para evaluar la correlación se utilizó el software R, librerías sp, maptools y stats (tabla 4, apéndice B: Tablas).

#### Resultados:

Quedaron seleccionadas las covariables, temperatura, temperatura por rangos, NDVI y Regiones naturales.

### 5.3. Análisis exploratorio de los datos

#### Metodología:

Se comenzó graficando la densidad de distribución de la variable resultado o dependiente (figura 9, apéndice A), la SMR para casos de hidatidosis en chicos de 0 a 14 años del área programa hidatidosis, Río Negro. Luego se utilizó una transformación logarítmica simple sobre esta variable, a fin de poder utilizar procedimientos autorregresivos, donde hay un supuesto subyacente de normalidad que los datos no parecen cumplir, ver figura 10, apéndice A.

También se realizaron los histogramas (figura 11, apéndice A) de SMR (ver capítulo 4, sección 4.1.) y  $\log(\text{SMR})$  con la curva de la distribución normal superpuesta (figura 12, apéndice A), a fin de comparar visualmente la normalidad de los datos.

Por otro lado se utilizaron *Scatterplot* para explorar la relación entre variable resultado y variables explicativas, temperatura media anual, NDVI y región ecológica PEAO. En otros *Scatterplot* se distinguieron los radios censales donde está presente o no la región ecológica PEAO y también presencia o ausencia de casos de hidatidosis.

#### Resultados:

A continuación, se observa en la figura 10 y 12 (apéndice A), que el  $\log(\text{SMR})$  se aproxima más a una distribución normal.

En cuanto a la relación entre variable respuesta y variables explicativas, se observó (figura 13, apéndice A) que a medida que aumenta la temperatura media anual de la superficie terrestre, aumenta el logaritmo del porcentaje de la SMR de hidatidosis. Mientras que cuando se evaluó la covariable en rangos (figura 14 y 15, apéndice A), la mayor correlación se dio con valores de 17,19 a 19,90 °C.

Pudo verse en la figura 14 y 15 (apéndice A) que los radios censales pertenecientes al rango de temperatura LST que va de 17.19 a 19.90 °C tienen mayor SMR o  $\log(\text{SMR})$ .

Cuando se fijó la temperatura media anual de la superficie terrestre como variable binaria, otorgándole valor 1 al rango de valores que va de 17.19 a 19.90 °C y 0 al resto, se observó (figura 16 y 17, apéndice A) mayor correlación en la media del rango fijado como 1.

Puede verse (figura 18, apéndice A) que a medida que disminuye el NDVI medio anual, aumenta el logaritmo del porcentaje de la SMR de hidatidosis.

Se observa (figura 19, apéndice A) que la media de la región ecológica PEAO se corresponde con un valor mayor del porcentaje de la SMR de hidatidosis, que la media del resto de las regiones ecológicas.

En la figura 20 (apéndice A), puede verse la relación entre la SMR y la temperatura media anual distinguida por rangos.

Puede verse (figura 21 y 22, apéndice A) que la región ecológica “Patagonia extra andina occidental” y la distribución de los casos de hidatidosis, coinciden con un rango de temperatura de 17,19 a 19,90 °C y con NDVI medio anual bajo.

En las Figuras 23 y 24 (apéndice A) se observa la relación entre SMR y temperatura y entre SMR y NDVI, mostrando la presencia o ausencia de casos de hidatidosis.

#### **5.4. Modelos Jerárquicos Bayesianos**

##### Metodología

Se desarrollaron modelos bayesianos de complejidad creciente, denominándose, de EF y de efectos mixtos (EM) (ver sección 4.3 a 4.7) con la incorporación de efectos aleatorios no correlacionados y efectos aleatorios correlacionados en el espacio.

##### 5.4.1. MODELO ESPACIAL DE EF (REGRESIÓN DE POISSON)

##### Metodología

Debido a que la variable respuesta (casos de hidatidosis en chicos de 0 a 14 años) es un recuento del número de eventos que ocurre en una población de un determinado tamaño, es una enfermedad que no se contagia persona a persona (si a través del ambiente y entorno contaminado) y es rara, se asume que los conteos siguen una distribución de Poisson, (Pfeiffer et al., 2008).

El conteo de la enfermedad observada en cada radio censal  $i$  asumió una distribución de Poisson con media  $u_i$  y ésta fue función de la cantidad de enfermedad esperada  $E_i$  (que depende del tamaño de la población) y del riesgo relativo  $RR_i$ , en ese radio censal (ver ecuaciones 1 y 2 en sección 4.1.2).

$$u_i = E_i RR_i \quad (3)$$

El riesgo relativo se parametrizo como función de las covariables  $x_{mi}$  definidas previamente (ver sección 5.2.2.).

$$\log(RR_i) = (\alpha + \beta x_i + \dots + \beta_m x_m) + \varepsilon_i \quad (4)$$

Entonces el número real de conteos de enfermedad en cada radio censal queda dado por (ver sección 4.3):

$$\log(u_i) = \log(E_i) + (\alpha + \beta x_i + \dots + \beta_m x_m) + \varepsilon_i \quad (5)$$

El término  $\varepsilon_i$  representa los residuos del riesgo relativo en el radio censal  $i$  después de ajustar por las covariables incluidas en el modelo. Se interpreta  $\varepsilon_i$  como residuos que reflejan la variabilidad entre áreas, debido a factores de riesgo desconocidos o no medidos.

El  $\log(E_i)$  es un factor de ajuste conocido como offset y representa el ajuste de acuerdo con áreas de diferentes tamaños de población.

Para determinar la influencia de las covariables sobre los casos de hidatidosis con los modelos de regresión de efectos fijos, primero se hizo una aproximación frecuentista (en R) y luego con el método bayesiano (usando OpenBUGS).

#### 5.4.1.1. REGRESIÓN CLÁSICA

##### Metodología

Se tomaron las covariables seleccionadas anteriormente y se procedió con un proceso de eliminación hacia atrás (*Backward elimination*), introduciendo todas las variables en la ecuación y luego excluyendo una tras otra al cumplir con el criterio de eliminación, que fue, la estimación de los coeficientes cercana a 0 y/o valor de significancia superior a 0.05 (ver tabla 5, 6 y 7, apéndice B).

El procedimiento termino cuando ya no quedaron en la ecuación variables que satisfacen el criterio de eliminación.

Para chequear por sobredispersión (ver sección 4.4.) se observó la relación entre la desviación residual y los grados de libertad provistos en el resumen, si la relación es mayor a 1 se puede decir que estamos en presencia de sobredispersión.

## Resultados:

Se puede observar (tabla 6, apéndice B) como la variable precipitación cumple con el criterio de eliminación (valor  $p > 0.05$ ), por lo que se excluye en el próximo modelo formulado.

Se puede ver (tabla 7, apéndice B) que en el segundo modelo no hay variables que cumplan el criterio de eliminación (valor  $p > 0.05$ ), por lo que el proceso de eliminación hacia atrás concluye aquí.

La relación entre la desviación residual y los grados de libertad provistos en el resumen (tabla 7, en apéndice B), es 3.45 (*residual deviance* = 172.57 / *degrees of freedom* = 50) siendo mayor a 1, lo que indica sobredispersión.

### 5.4.1.2. REGRESIÓN BAYESIANA

#### Metodología

Para especificar, ejecutar y evaluar los modelos de regresión Bayesiana utilizamos los softwares OpenBUGS y R con sus paquetes R2OpenBUGS y CODA (ver apéndice C). El software OpenBUGS se llamó desde dentro de R y se crearon R objetos que se integraron perfectamente con varias herramientas espaciales.

Creación del modelo, datos y parámetros (tabla 8, apéndice B).

Se establecieron los datos y variables para ejecutar el modelo (tabla 9, apéndice B). Se crearon objetos para los datos, definiciones de variables y los valores iniciales para la función bugs () utilizando el paquete R2WinBUGS.

Se puede observar en el apéndice D, la evaluación de la salida del muestreo de Gibbs (ver también sección 4.2) con CODA.

Como buena práctica para evaluar el rendimiento del modelo se usaron diferentes valores iniciales, lo que se hizo corriendo varias cadenas.

También para chequear que la convergencia ha sido adquirida, como regla general, se observó que  $MC\ error[i]/SD[i]$  sea menor que 0.05, siendo la variable MC error, provista por la salida de WinBUGS u OpenBUGS y es el error estándar de Monte - Carlo de la media, que es igual a  $\sigma/N^{0.5}$ .

Finalmente se compararon los coeficientes de regresión de los modelos de EF calculados previamente con estadística frecuentista (en el software R) y los calculados con estadística Bayesiana (en OpenBUGS).

## Resultados

Se muestran los resultados (tabla 10, en apéndice B) de la comparación de los coeficientes de regresión de los modelos de EF computados usando estadística frecuentista y estadística Bayesiana. Los dos métodos produjeron resultados similares.

La siguiente tabla (tabla 11, apéndice B), contiene un resumen de estadísticos del modelo de EF Bayesiano que se denominó MOD\_1.

### **INTERPRETACIÓN DE LOS COEFICIENTES DE REGRESIÓN**

Basándonos en la tabla 11 (apéndice B), la interpretación de los coeficientes de regresión del modelo puede expresarse de la siguiente forma, por cada unidad de incremento en la variable explicativa (covariable) hay un aumento (+) o disminución (-) en el número medio de la variable explicada por un factor multiplicativo de  $\exp(\beta)$  o expresado en %  $(\exp(\beta)-1) * 100$ . Dicho de otra manera y para cada covariable (considerando que las demás covariables permanecen constantes), por unidad de incremento en la temperatura media anual, el riesgo relativo de hidatidosis disminuye 0.27 veces ( RR 0.73, 95% IC 0.66 - 0.82) y por 0.1 de incremento en el NDVI, el riesgo relativo de hidatidosis disminuye 1 vez (RR 0.00, 95% IC 0.00 - 0.00), por lo que zonas con valores elevados de temperatura y NDVI, actúan como factor de protección, en menor y mayor medida respectivamente. En el caso de covariables binarias un cambio de 0 a 1, produce un aumento (+) o disminución (-) en la media de la variable explicada por un factor de  $\exp(\beta)$ . De otra forma, para la variable, región ecológica PEAO, puede decirse que hay 33.38 veces más chances de enfermar de hidatidosis (RR 33.38, 95% IC 15.30 – 76.02) si el radio censal está ubicado en esta región ecológica (1) en contraste con estar ubicado en cualquier otra región ecológica (0).

#### 5.4.2. RESIDUOS Y AUTOCORRELACIÓN

### Metodología

Se trabajó en la diferencia (residuo), entre el número observado en cada región y el número predicho por el modelo, debido a que un modelo provee un medio para predecir resultados en base a un set de variables explicativas y el modelo desarrollado podría ser usado para predecir el número de casos de hidatidosis diagnosticado en cada región en base a la temperatura media anual, NDVI y región ecológica PEAO.

Se evaluó la autocorrelación en los residuos como guía para decidir la introducción de efectos aleatorios estructurados (ver sección 4.6) en el modelo, aunque la solo mejora

en el DIC (ver sección 4.9.1) puede justificar esto. Primero se evaluó la autocorrelación en los residuos del modelo de EF y luego en el modelo con efectos aleatorios incorporados a fin de corroborar si existe disminución en la autocorrelación.

Para cuantificar la autocorrelación espacial en áreas, varias medidas son disponibles. Aquí se utilizó el estadístico Moran's I (MORAN, 1950), ver tabla 12 (apéndice B), que provee una medida de la similitud de una variable entre áreas definidas como cercanas en el espacio (sobre la base de una matriz de adyacencia espacial, definida por el analista). Moran's I se aproxima a una distribución normal y su valor esperado es  $-1/(N-1)$ , siendo N igual al número de áreas dentro de una región de estudio. El valor esperado de Moran's I es negativo y es función del número de áreas dentro de la región de estudio y se acerca a 0 cuando el número de áreas se incrementa. Aunque Moran's I generalmente reside entre  $+1$  y  $-1$  no está obligado a estos límites (Pfeiffer et al., 2008). Un Moran's I cero indica la hipótesis nula de no agrupación, un Moran's I positivo indica autocorrelación espacial positiva (agrupación de áreas de valores similares), un coeficiente negativo indica autocorrelación espacial negativa (áreas vecinas que tienen valores de atributo diferentes).

Se utilizaron indicadores locales de asociación espacial (LISA) para indicar la presencia o ausencia de agrupaciones espaciales significativas o valores atípicos para cada ubicación. Estos se basan en un enfoque de aleatorización para generar una distribución de referencia espacialmente aleatoria para evaluar la significación estadística. El estadístico de Moran local implementado en GeoDa (GeoDa Center for Geospatial Analysis and Computation, s.f.) es un caso especial de un LISA (<https://geodacenter.asu.edu/node/390#lisa2>). El promedio de las estadísticas de Moran local es proporcional al valor del Moran Global I. Mapas LISA son particularmente útiles para evaluar la hipótesis de aleatoriedad espacial y para identificar los puntos calientes locales.

## Resultados

A continuación, pueden observarse los mapas de Casos, Predichos y residuos (figura 25, apéndice A). La diferencia entre Observados y predichos varía de -16.4 (a favor de los predichos) a 9.2 (a favor de los observados).

Se puede observar (tabla 13, apéndice B) que el estadístico de Moran's I es negativo -0.16 (autocorrelación espacial negativa: áreas vecinas que tienen valores de atributo diferentes) y no significativo (valor  $p = 0.96$ ) a un valor de  $\alpha = 0.05$ . El valor esperado es  $-1/(54-1)$  igual a -0.0188.

Hay cuatro outlier marcados en el Moran's I *scatterplot* (figura 26, apéndice A), 4, 15, 37 y 43, que pueden observarse en la figura 27 (apéndice A) también. El radio censal 4 es un área con valores de residuos altos (más observados que predichos) rodeado por radios censales con valores de residuos bajos (menor diferencia entre observados y

predichos), presentando autocorrelación espacial negativa. El radio censal 37 tiene valor de residuo alto (más alto que el área 4) y está rodeado por radios censales con valores altos y medios. En el caso del radio censal 15, se observa un valor de residuo bastante negativo (más predichos que observados) y está rodeado por radios censales con valores de residuos altos y medios altos. Por último, el radio censal 43 no posee valores de residuos tan negativos como el área 15, pero se autocorrelaciona negativamente con valores altos.

En la figura 28 (apéndice A) pueden observarse los LISA donde el radio censal 4 es un valor alto rodeado de valores bajos y el radio censal 43 es un valor bajo rodeado de valores altos, ambos significativos a un nivel  $\alpha$  de 0.05.

#### 5.4.3. SOBREDISPERSIÓN

##### Metodología

Para tratar con la sobredispersión (ver sección 4.4) o extra-variación se extendió el modelo por medio de un término lineal o predictor no lineal para incluir un efecto aleatorio adicional (modelo log-normal, ver sección 4.5).

#### 5.4.4. MODELOS ESPACIALES DE EFECTOS MIXTOS (MODELO LOG-NORMAL)

##### Metodología

Para cuantificar la influencia de EF (ver sección 4.7) en la enfermedad para cada área o región de estudio, e identificar áreas donde hay conteos de enfermedad elevados luego de que la influencia de los EF han sido evaluados, se utilizó un modelo mixto Bayesiano completo donde efectos aleatorios correlacionados a nivel de área son determinados (Toledano, Best, Wakefield y Elliott, 2001; Jarup et al. 2002).

En este método se incluyen efectos aleatorios área-específicos para evaluar características espaciales dentro del dato y en vez de producir un coeficiente de autocorrelación espacial (global), efectos espaciales aleatorios son determinados para cada área. Esto sirve para identificar áreas donde hay un exceso de enfermedad que no es representado por EF incluidos en el modelo y en términos prácticos focalizar estas áreas para futuros esfuerzos de vigilancia epidemiológica.

Los problemas metodológicos sobre este tema pueden verse en Besag et al. (1991), Clayton y Bernardinelli (1992), Clayton y Kador (1987), Cressie (1992), Pickle et al. (1996), Waller et al. (1997), Xia et al. (1997) y Conlon y Louis (1999).

Se formuló el siguiente modelo jerárquico bayesiano espacial para conteo de enfermedades. El número medio de eventos de enfermedad en un área se parametrizo (se explicó) en términos de una serie de covariables área-específicas (ver ecuación 5 en sección 5.4.1).

En un contexto bayesiano una distribución a priori no-informativa se asumió para el intercepto  $\alpha$  y para cada uno de los coeficientes  $\beta_m$ . El exponente de  $\varepsilon_i$  representa los residuos del riesgo relativo en el área  $i$  luego de ajustar por las covariables incluidas en el modelo. Se interpreta a  $\varepsilon_i$  como el reflejo de la variabilidad de los residuos entre áreas debido a factores de riesgo desconocidos o no medurados.

Debido a que los factores de riesgo no medidos, generalmente varían suavemente en el espacio y a su vez inducen correlación espacial entre el conteo de enfermedad observado en áreas cercanas, se evaluó la presencia de correlación espacial, asumiendo que los efectos aleatorios están compuesto por dos tipos: un componente estructurado (espacialmente correlacionado) y un componente no estructurado.

$$\log(u_i) = \log(E_i) + (\alpha + \beta x_i + \dots + \beta_m x_m) + v_i + u_i + \varepsilon_i \quad (6)$$

En la ecuación precedente,  $u_i$  representa el efecto aleatorio estructurado (espacialmente correlacionado) y  $v_i$  representa el efecto aleatorio no estructurado para cada área  $i$ . Se asumió una distribución a priori espacialmente estructurada para el componente estructurado de los efectos aleatorios. Varias elecciones existen pero se trabajó con la más popular, un caso especial de Modelo CAR (Besag et al., 1991; Besag y Kooperberg, 1995). Este modela el logaritmo del riesgo relativo en el área  $i$  condicional al riesgo en todas las otras áreas siendo normalmente distribuido alrededor del peso de la media del logaritmo de riesgo relativo en las áreas remanentes, con la suma de los pesos siendo inversamente proporcional a la varianza  $\sigma^2$ . Los efectos aleatorios no estructurados  $v_i$  se parametrizan siendo normalmente distribuidos con media cero y varianza  $\tau^2$ .

Se asignan distribuciones a priori (hyperpriors) que asumen una distribución gamma con constante  $a$  fija (el parámetro “*shape*”) y  $b$  (el parámetro inverso “*scale*”) tal que la media de la distribución es igual  $a/b$  y la varianza igual a  $a/b^2$ .

## EFFECTOS ALEATORIOS ESTRUCTURADOS

Como áreas cercanas en el espacio tienden a tener riesgos relativos similares, se asumió que para esta clase de modelos, la distribución condicional del riesgo relativo en la región  $i$ , depende de valores de RR en los vecinos.

Se especificó una distribución a priori condicional autorregresiva (CAR) usando el vector `CAR.normal` en OpenBUGS:

$$u[1:N] \sim \text{CAR.normal}(\text{adj}[], \text{weights}[], \text{nneigh}[], \text{tau.u})$$

Donde `adj []` es un vector conteniendo una lista con la identificación de las áreas adyacentes para cada área (esta es una sencilla representación de la matriz de adyacencia espacial completa para la región de estudio); la longitud de `adj []` es igual a la suma de los  $n$  vecinos. `Weights []` es un vector igual de largo que el anterior, dando pesos sin normalizar, asociados a cada par de áreas. Para el modelo CAR, se toma  $C_{ij} = 1$  (equivalente  $W_{ij} = 1/n$  si áreas  $i$  y  $j$  son vecinos, y 0 de otro modo), teniendo un vector de unos para `weights[]`. Se fijaron todos los pesos iguales a 1, lo que es una elección común (Besag et al., 1991). `nneigh []` es un vector de largo  $N$  (el número total de áreas) conteniendo el número de vecinos  $n_i$  para cada área. `tau.u` es un argumento escalar (un valor simple) representando la precisión (la inversa de la varianza) de la distribución a priori de CAR. La distribución especificada para `tau.u` determinara la fuerza de los efectos aleatorios estructurados en el modelo.

Pequeños valores de `tau.u` (una gran varianza en la distribución a priori de CAR) fuerzan a los efectos aleatorios espaciales a ser diferentes – una pequeña cantidad de suavizado se aplica al dato. Grandes valores de `tau.u` (una pequeña varianza en la distribución a priori de CAR) fuerza a los efectos aleatorios espaciales a ser similares – una gran cantidad de suavizado espacial es aplicada al dato. Si se quiere inducir una pequeña cantidad de suavizado espacial, se parametriza `tau.u` con un pequeño *shape* parámetro. Si se quiere inducir una gran cantidad de suavizado espacial, se parametriza `tau.u` con un *shape* parámetro grande. Un parámetro *scale* inverso grande hará la varianza de `tau.u` pequeña (se debe confiar acerca del grado de suavizado espacial que se está aplicando). Un parámetros *scale* inversa pequeño hará la varianza de `tau.u` grande (no se debe confiar acerca del grado de suavizado espacial que se está aplicando).

Aquí se asume una distribución uniforme impropia  $(-\infty, +\infty)$  para el intercepto global (Alpha), que es parametrizado como `dflat()` en el lenguaje de WinBUGS u OpenBUGS, debido a que la distribución a priori `CAR.normal` incluye una restricción *sum-to-zero*, que significa que hay que incluir un intercepto separado en el modelo, que pueda dar una prior impropia (localización invariante) en orden de que el modelo sea equivalente a una parametrización sin restricciones del modelo CAR sin intercepto global.

## EFECTOS ALEATORIOS NO ESTRUCTURADOS

La distribución para los efectos aleatorios no estructurados se especificó como normal  $(0, \text{tau.v})$  en OpenBUGS. Aquí, `tau.v` es un argumento escalar representando la precisión de los efectos aleatorios no estructurados.

Se aplicaron distribuciones a priori difusas a los efectos aleatorios no estructurados y una gran varianza se aplicó a la distribución normal, siendo buena práctica para evaluar la sensibilidad y para asumir distribuciones a priori de los resultados finales obtenidos.

#### 5.4.4.1. REGRESIÓN BAYESIANA DE EFECTOS MIXTOS.

##### Metodología

Para especificar, ejecutar y evaluar los modelos de regresión Bayesiana de efectos Mixtos, utilizamos nuevamente los software OpenBUGS y R con sus paquetes R2OpenBUGS y CODA.

Creación del modelo, datos y parámetros (Tabla 14 y 15, apéndice B).

Con el fin de poder ver los aportes de cada variable incluida en los modelos, se realizó la estandarización y centrado de estas (tabla 16, apéndice B). Para ello se utilizó la función `rescale()` del `package(arm)` del software R, la que estandariza una variable al centrarla y dividirla por 2 sd's con excepción de variables binarias, permitiendo de este modo su comparación (Gelman, 2008).

Luego de estandarizar y centrar las covariables se siguió el procedimiento como en la Tabla 15 (apéndice B), definiendo los demás datos como el número de regiones, los casos y los esperados.

##### Resultados

Se puede ver en la tabla 17 (apéndice B) los resultados de los modelos de EM (MOD\_2: con efectos aleatorios no estructurados y MOD\_3: con ambos efectos aleatorios, estructurados y no estructurados) comparados con el modelo de EF (MOD\_1).

En la tabla 18 (apéndice B) se puede observar la bondad de ajuste (ver sección 4.9.1) para los diferentes modelos. Vemos que los ítems MSPE, desviación y DIC son menores para los modelos de EM (MOD\_2 y MOD\_3) lo que indica que tienen mejor bondad de ajuste, sobre todo MOD\_3 que posee el DIC más pequeño. En cuanto al número de parámetros estimado (pD) es obvio el aumento a medida que los modelos incrementan su complejidad.

Se puede ver en los modelos con las covariables estandarizadas y centradas (tabla 19, apéndice B), que la influencia de los coeficientes de regresión de las covariables difieren en menor medida que antes de estandarizar y centrar, siendo los aportes de las covariables temperatura media anual y NDVI más parecidas (por ejemplo en MOD\_3 es

de -2.94 y -3.85 respectivamente) y la covariable PEAO con influencia positiva se mantiene similar como es de esperar (2.29 para MOD\_3).

Puede observarse en la tabla 20 (apéndice B), el modelo de EF y EM para la temperatura media anual de la superficie terrestre en rangos. Aquí se observó el riesgo relativo de hidatidosis en los radios censales que están ubicados en un rango de temperatura de 17.19 a 19.90 °C en contraste con los que están fuera de este rango. La verdadera influencia de este rango se ve en el modelo de EM, cuando parte del aporte de la temperatura es realmente debido a efectos aleatorios, pasando a ser el RR de 93.78 para el modelo de EF a 11.23 para el modelo de EM.

#### 5.4.4.2. ANÁLISIS DE LA PREDICCIÓN DEL MODELO.

##### Metodología

Se evaluaron los mapas resultantes del modelo de EM con efectos aleatorios (estructurados y no estructurados) incorporados y se compararon los mapas de casos, predichos y residuos, para ver la predicción del modelo.

##### Resultados

Se observan los mapas resultantes del modelo de EM con efectos aleatorios (estructurados y no estructurados) incorporados.

En la figura 29 (apéndice A) se muestran los mapas de casos (vistos previamente en la figura 25, apéndice A), el mapa de casos predichos por el modelo MOD\_3 y el resultante de la diferencia de los anteriores (residuos). Puede observarse que el mapa de casos predichos es ahora mucho más parecido al mapa de casos observados, luego de la incorporación de efectos aleatorios al modelo de EF (MOD\_1), lo que se denota también por la reducción en el rango de los residuos, que iba de -16.4 a 9.2 en MOD\_1 y se redujo a un rango de -0.9 a 1.5, en MOD\_3.

#### **INFORMACIÓN ADICIONAL Y EFECTOS ALEATORIOS DEL MODELO DE EM**

Aquí se tienen en cuenta dos cuestiones, primero la información adicional que se obtiene del modelo de EM y segundo, la interpretación de los efectos aleatorios del modelo.

En cuanto a la primera cuestión, luego de evaluar los efectos aleatorios espaciales y no espaciales, la influencia de la región ecológica PEAO ha disminuido (ver RR en tabla 17, apéndice B), pasando el RR de 33.38 en el modelo de EF a 10.04 en el modelo de

EM (MOD\_3). Esto implica que hay influencias espacialmente correlacionadas y no correlacionadas sobre el número de casos de hidatidosis observados.

Para la interpretación de los efectos aleatorios del modelo (ver heterogeneidad estructurada y no estructurada en tabla 17, apéndice B), podemos decir, que la desviación estándar (SD) de los efectos aleatorios estructurados es de 0.22 y la SD de los no estructurados es de 1.80, indicando que los efectos no estructurados son 8 veces más variables que los estructurados. Esto sugiere que los efectos aleatorios no estructurados son importantes en algunas áreas de la región de estudio y en otras no. El efecto aleatorio estructurado, en contraste, es relativamente uniforme para todas las áreas. Puede también observarse el parámetro Alfa en MOD\_3, el que explica la relación entre los dos componentes de la heterogeneidad, siendo mayor el componente estructurado a medida que Alfa se acerca a 1, y menor al acercarse a 0. En este caso es 0.09, o sea, representa solamente cerca del 10% de los efectos aleatorios.

#### 5.4.4.3. ANÁLISIS Y DESCOMPOSICIÓN DEL RR

##### Metodología

Se graficaron los intervalos de credibilidad del 95% para la estimación del riesgo relativo de los efectos de las covariables sobre el número de casos de hidatidosis, con el fin de comparar estos para los modelos de EF y EM.

Por otro lado se observaron y compararon los mapas de la SMR y el RR y los residuos del RR para el modelo de EM MOD\_3. El residuo del RR se obtuvo por substracción de las covariables.

Además para ayudar a la interpretación se representó en mapas la descomposición del RR fijado en sus tres componentes (Lawson, Browne y Rodeiro, 2003), variables explicativas ( $\exp(\beta x_i)$ ), heterogeneidad no estructurada ( $\exp(v_i)$ ) y efectos aleatorios estructurados ( $\exp(u_i)$ ).

##### Resultados

Las siguientes figuras (30, 31 y 32, apéndice A) muestran el intervalo de credibilidad del 95% para la estimación del riesgo relativo de los efectos de las covariables sobre el número de casos de hidatidosis. Se puede observar que en el caso de la temperatura media anual y NDVI, el valor medio del RR en los modelos mixtos se mantiene prácticamente igual y sus intervalos de credibilidad siguen estando por debajo de 1, lo que reafirma que el aumento del valor de estas variables influye en forma negativa sobre el número medio de casos de hidatidosis en nuestros modelos. Mientras que en el caso de la región ecológica PEAO, la media del RR disminuye, implicando que parte del

riesgo que fue atribuido a la influencia de esta región en el modelo de efectos fijos, se desvía ahora hacia los efectos aleatorios (espaciales y no espaciales) no medidos en los modelos de EM, aunque el intervalo de credibilidad se mantiene sobre 1, ratificando que el RR aumenta cuando un radio censal está sobre esta región ecológica.

Una serie de mapas vistos en la figura 33 (apéndice A) corresponden a la SMR, el RR y los residuos del RR. El primer mapa (superior) muestra la SMR, el segundo mapa (medio) el RR del modelo de EM con covariables incorporadas y el tercer mapa (inferior) muestra los residuos del RR, obtenido por sustracción de las covariables. Se puede observar que las áreas con riesgo relativo entre 25 a 100 del mapa superior (áreas ubicadas al sur y desplazadas levemente, desde el centro hacia el oeste), muestran valores menores a 25 en el tercer mapa. Confirmando de esta manera, el aporte de las covariables al incremento en el RR en estas zonas. Mientras que en las áreas ubicadas al Noreste y algunas áreas del oeste del mapa de residuos de RR, se observa un aumento de RR, debido a la sustracción de las covariables que influyen negativamente en el RR en estas zonas.

En la figura 34 (apéndice A) se puede observar la descomposición del RR fijado en sus tres componentes (Lawson et al., 2003), debido a variables explicativas ( $\exp(\beta x_i)$ ), a heterogeneidad no estructurada ( $\exp(v_i)$ ) y a efectos aleatorios estructurados ( $\exp(u_i)$ ).

Comparando los mapas de la figura 34 puede verse que las covariables influyen en el RR de las regiones ubicadas al sur y centro, y al sur y centro oeste del área de estudio. Mientras que la región (radio censal 4) con  $RR > 100$  en el norte y centro del mapa de RR de heterogeneidad, parece ser explicado por efectos aleatorios no estructurados desconocidos. Por último puede observarse en el mapa de RR debido a efectos aleatorios estructurados, que el RR en la mayoría de las regiones es homogéneo e inferior a 25.

#### 5.4.4.4. ANÁLISIS DE SENSIBILIDAD PARA DISTRIBUCIÓN A PRIORI

##### Metodología:

El análisis de sensibilidad se realizó a fin de corroborar que no haya cambios en los resultados del análisis ante información previa diferente. Las distribuciones a priori para los coeficientes de las covariables han sido especificadas para ser normalmente distribuidas con media cero y precisión  $1 \times 10^{-5}$  ( $\text{dnorm}(0.0, 1.0E-5)$ ). La inversa de la precisión (que es la varianza) fue igual a un número grande, esto significa que no se es específica acerca de la creencia previa sobre el valor de  $\beta_1$ ,  $\beta_2$  y  $\beta_3$  (se fija una distribución a priori “vaga” o “difusa”, ver sección 4.3 “Modelo lineal Bayesiano”), al hacer esto, se deja que los datos observados influyan en mayor medida la determinación de la distribución a posteriori.

En el enfoque bayesiano de mapeo de enfermedades, la elección de la distribución hyperprior de los parámetros de precisión es una cuestión clave. Para investigar la influencia de la elección de la hyperprior en estimaciones de riesgo relativo, se llevo a cabo un análisis de sensibilidad para diferentes opciones de distribuciones a priori o parámetros. Si los datos son numerosos, los datos dominan la distribución a priori y por lo tanto importa menos que se elija como valores de los hiperparámetros. Si los datos son escasos, la elección de una combinación adecuada de hiperparámetros es importante. En nuestro caso se utilizo una especificación Gamma(0.5, 0.0005) común (Thomas et al., 2004), para la distribución hyperprior del parámetro inversa de la varianza de los efectos aleatorios.

Se hicieron modificaciones de las distribuciones a priori y se volvieron a calcular las cantidades a posteriori de interés, luego se registro si estas imponen un impacto práctico sobre interpretaciones o decisiones. Para tal fin se adaptó una distribución a priori informativa Gamma(0.1, 0.1), y otras, Gamma(0.01, 0.01) y Gamma(0.001, 0.001).

#### Resultados:

Para el análisis de sensibilidad, en la tabla 21 (apéndice B) puede verse el resultado para la especificación Gamma(0.5, 0.0005) utilizada y su comparación con una distribución a priori informativa Gamma(0.1, 0.1), y otras, Gamma(0.01, 0.01) y Gamma(0.001, 0.001). Las cuatro distribuciones dieron resultados similares para todos los parámetros, lo que indica que los resultados son robustos a cambios en la información previa. Además diferencias en DIC menores que 5 no son de importancia cuando los modelos hacen inferencias similares (MRC Biostatistics Unit, 1996 - 2004).

#### 5.4.4.5. ANÁLISIS DE LA AUTOCORRELACIÓN

##### Metodología

Finalmente se evalúa la autocorrelación de los residuos en el modelo de EM con efectos aleatorios incorporados, para ver si el índice de Moran's I se aproxima al valor esperado.

El procedimiento es igual al descrito previamente (sección 5.4.2) para el modelo de efectos fijos.

##### Resultados

Podemos ver en la tabla 22 y en la figura 35 (apéndice B y A respectivamente) el resultado de la evaluación de autocorrelación de los residuos en el modelo de EM con

los efectos aleatorios incorporados, observándose ahora que el estadístico Moran'I se acerca más al valor esperado.

Podemos ver en la figura 36 (apéndice A) los radios censales que corresponden a valores atípicos. El radio censal 4 presenta valores de residuos altos rodeado por valores de residuos bajos. Los radios censales 2 y 7 presentan valores de residuos cercanos a cero y están rodeados por otros radios censales con valores similares. El radio censal 29 presenta un valor de residuo alto rodeado por valores relativamente bajos.

En la figura 37 (apéndice A) pueden observarse los LISA donde hay 5 radios censales con residuos con valores bajos rodeados por valores altos significativos a un nivel  $\alpha$  de 0.05.

#### 5.4.4.6. ANÁLISIS DE RESIDUOS

##### Metodología

Se procedió al análisis de los residuos mediante el gráfico de observados y predichos (figura 38, apéndice A), y el gráfico de residuos estandarizados y predichos (figura 39, apéndice A), que muestran el grado de correlación y exactitud de la predicción.

##### Resultados

En general cuando un modelo es adecuado y ajusta bien pueden esperarse una serie de características en los residuos. En primer lugar debiesen ser simétricos y centrados alrededor de cero.

En la figura 38 (apéndice A) puede verse el gráfico de observados vs. predichos, donde hay una fuerte correlación entre las predicciones del modelo y sus resultados reales.

Luego observamos el gráfico de residuos estandarizados vs. predichos (figura 39, apéndice A), en este gráfico cada punto es de un caso, donde la predicción hecha por el modelo está representada en el eje x, y la exactitud de la predicción en el eje y. La distancia a la línea 0 es lo mal que la predicción es para ese valor. Valores positivos para el residuo (en el eje y) significa que la predicción es demasiado baja, y valores negativos significa la predicción es demasiado alta, 0 significa que la conjetura es exacta.

De manera ideal los residuos tendrían que estar simétricamente distribuidos, tendiendo a agruparse hacia el centro de la trama, agrupados alrededor de los dígitos cercanos a 0 del eje y, y en general no presentar patrones claros.

En nuestro caso se observan algunos puntos que se alejan por encima de 2 en el eje y. También puede observarse una pequeña estructura entre 0 y 1 en el eje x, que si se ve el gráfico de observados vs. predichos, parece ser que valores 0 observados, se predicen con un valor superior a este. Por lo demás creemos son aceptables.

---

## CONCLUSIONES.

### CAPÍTULO 6:

#### 6. CONCLUSIONES

Se logró identificar áreas de prevalencia elevada y explicar estas en términos de covariables ambientales y climáticas, encontrándose correlación entre áreas con tasa de hidatidosis elevadas y valores bajos de NDVI y de LST, y también encontrándose correlación con la región PEAO en contraste con otras regiones ecológicas y con el rango de LST de 17,19 a 19,90 °C comparado con otros rangos.

Luego de evaluar los factores ambientales y climáticos se determinó el exceso de enfermedad en cada área, representando en mapas el riesgo relativo para EF y efectos aleatorios. En general los EF (covariables) explican las tasas elevadas de hidatidosis en la mayoría de las áreas, salvo un área ubicada al centro - norte provincial, donde la tasa elevada puede explicarse por efectos aleatorios no estructurados en el espacio.

La correlación encontrada en nuestros resultados entre regiones con tasas altas de hidatidosis y temperaturas bajas (modelos de tres covariables, MOD\_1 a MOD\_3) y medias (modelo con la temperatura por rangos como covariable), es consistente con el hecho de que el clima cálido sea un factor que afecte la vitalidad de los huevos de *Echinococcus granulosus* como se describe en Lawson y Gemmell, 1983 y Guarnera, 2009, que también describen el clima seco como influencia negativa sobre los huevos, cuestión que no coincidiría con nuestros hallazgos, si se evalúa indirectamente la humedad a través del NDVI, dado que en las regiones con menor NDVI se encuentran las tasas más elevadas de enfermedad. Todo esto teniendo en cuenta, que a mayor oferta de huevos, mayor tasa de enfermedad.

La correlación de tasas elevadas de hidatidosis con la región ecológica PEAO, indica que características biofísicas, socioeconómicas y productivas, propias de esta región, por si solas o en conjunto, influyen sobre la distribución geográfica de la enfermedad. Además la correlación encontrada, concuerda con lo mencionado por Larrieu et al. 1986 y 2000, sobre las áreas geográficas de mayor riesgo, sobre todo para la zona de Comallo, Pilcaniyeu, Ñorquinco e Ingeniero Jacobacci.

El modelo desarrollado podría ser usado para predecir la tasa de enfermedad en cada región en base a la temperatura media anual de la superficie terrestre, NDVI y región ecológica PEAO para las zonas mencionadas en el párrafo previo y para periodos similares al del estudio (10 años).

Algunos de los problemas que podemos mencionar de este estudio, pueden ser, los radios censales que toman formas arbitrarias y no la que deseáramos, pero generalmente, éste es un problema asociado al enfoque de áreas pequeñas (ver cita bibliográfica en la sección 1.1). El sesgo ecológico puede ser un problema, cuando las asociaciones establecidas se extrapolan a una escala menor a la del estudio. Otro defecto, puede ser el realizar promedios de las variables dentro de cada radio censal, esto puede llevar a la pérdida de variabilidad en los datos. Por último la relación que puede establecerse entre la variable dependiente (SMR de hidatidosis) y variables explicativas es indirecta, pero basada en el hecho de que a mayor vitalidad de los huevos de *Echinococcus granulosus*, mayor carga ambiental y mayor tasa de enfermedad.

Por otro lado creemos fue importante el hecho de realizar el cálculo de la SMR, solo con datos de casos y población rural para cada radio censal, ya que la hidatidosis es una enfermedad mayormente rural en nuestra provincia y tener en cuenta la población urbana distorsiona las tasas.

En futuras investigaciones podrían estudiarse con los mismos modelos la relación entre corderos infestados y/o suelo contaminado con huevos de *Echinococcus granulosus* (como variable dependiente) y las variables explicativas estudiadas en esta tesis, de forma de establecer una asociación más directa. Otro punto a estudiar es, cuales son los efectos aleatorios no correlacionados que influyen sobre la tasa elevada de hidatidosis en el radio censal 4, ubicado al centro - norte provincial. Quedaría también por dilucidar si existe alguna otra variable presente en la región ecológica PEAO, que influya por si sola sobre la tasa de enfermedad, como por ejemplo el sistema productivo. Por último, si se cuenta con casos y controles georreferenciados pueden desarrollarse modelos binomiales que evalúen las correlaciones encontradas a una escala geográfica menor.

## REFERENCIAS.

- Ammann, R., Eckert, J. (1995). Clinical diagnosis and treatment of echinococcosis in humans. In R. Thompson, & J. Lymbery. *The biology of Echinococcus granulosus and hydatid disease* (pp. 411-477). London: George Allen and Unwin.
- Arezo, M. A. (2012). Uso de MODIS en estudios relacionados con enfermedades transmitidas por vectores "Mosquitos" y su contribucion en la eleccion de sitios optimos de colocacion de trampas en la region del Veneto. Instituto de Altos Estudios Espaciales "Mario Gulich" - CONAE. Recuperado de [http://aulavirtual.ig.conae.gov.ar/moodle/pluginfile.php/513/mod\\_page/content/71/Informe\\_Arezo.pdf](http://aulavirtual.ig.conae.gov.ar/moodle/pluginfile.php/513/mod_page/content/71/Informe_Arezo.pdf).
- Atkinson, J. A. M., Gray, D. J., Clements, A. C., Barnes, T. S., McManus, D. P., & Yang, Y. R. (2013). Environmental changes impacting Echinococcus transmission: research to support predictive surveillance and control. *Global change biology*, 19(3), 677-688.
- Atkinson, P. M., & Graham, A. J. (2006). Issues of scale and uncertainty in the global remote sensing of disease. *Advances in parasitology*, 62, 79-118.
- Baddeley, A., & Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns. *Australian & New Zealand Journal of Statistics*, 42(3), 283-322.
- Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Statistics*. New York: Chapman&Hall/CRC.
- Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

- Berke, O. (2004). Exploratory disease mapping: kriging the spatial risk function from regional count data. *International Journal of Health Geographics*, 3(1), 18.
- Bernardinelli, L., Clayton, D., & Montomoli, C. (1995). Bayesian estimates of disease maps: how important are priors?. *Statistics in medicine*, 14(21-22), 2411-2431.
- Bernardinelli, L., Pascutto, C., Montomoli, Komakec, J. & Gilks, W. (1999). Ecological regression with errors in covariates. In A. B. Lawson, D. Boehning, E. Lasaffree, A. Biggeri, J. F. Viel & R. Bertolline. *Disease Mapping and Risk Assessment for Public Health* (pp.329-346). Chichester: John Wiley & Sons, Ltd.
- Bernardinelli, L., Pascutto, C., Montomoli, & Gilks, W. (2000). Investigating the genetic association between diabetes and malaria: an application of Bayesian ecoecological regression models with errors in covariates. In P. Elliott, J. Wakefield, N. G. Best, and D. Briggs. *Spatial Epidemiology. Methods and Applications*. Oxford: Oxford University Press.
- Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1–20.
- Besag, J., & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4), 733-746.
- Best, N., Waller, L., Thomas, A., Conlon, E., Arnold, R.(1999). Bayesian models for spatially correlated disease and exposure data. In J. Bernardo, J. Berger, A. Dawid, A. Smith. *Bayesian Statistics 6: Proceedings of the Sixth Valencia Meeting on Bayesian Statistics* (pp. 131– 156). Oxford : Oxford University Press.

- Biggeri, A., Divino, F., Frigessi, Lawson A.B., Boehning, D., Lesaffree, E., and Viel, J.F. (1999). Introduction to spatial models in ecological analyses. In A. B. Lawson, D. Boehning, E. Lasaffree, A. Biggeri, J. F. Viel, and R. Bertollini. *Disease Mapping and Risk Assessment for Public Health* (pp.181-191). Chichester: John Wiley & Sons, Ltd.
- Breslow, N. E., & Day, N. E. (1993). Statistical methods in cancer research. Vol II. The design and analysis of cohort studies. Lyon: International Agency for Research on Cancer, 1987. *IARC sci publ*, (82), 96-7.
- Britch, S. C., Linthicum, K. J., Anyamba, A., Tucker, C. J., Pak, E. W., Maloney, F. A., ... Miller, M. (2008). Satellite vegetation index data as a tool to forecast population dynamics of medically important mosquitoes at military installations in the continental United States. *Military Medicine*, 173(7), 677–683.
- Brooks, SP. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434-455.
- Budke, C. M., Deplazes, P., & Torgerson, P. R. (2006). Global socioeconomic impact of cystic echinococcosis. *Emerg Infect Dis*, 12(2), 296-303.
- Bustos, O. H., Frery, A.C., Lamfri, M.A. & Scavuzzo, C.M. (2004). Técnicas Estadísticas en Teledetección Espacial. Instituto de Altos Estudios Espaciales ‘Mario Gulich’, CONAE.
- Cabrera, P. A., Haran, G., Benavidez, U., Valledor, S., Perera, G., Lloyd, S., ... & Carballo, M. (1995). Transmission dynamics of *Echinococcus granulosus*, *Taenia hydatigena* and *Taenia ovis* in sheep in Uruguay. *International journal for parasitology*, 25(7), 807-813.
- Cabrera, P. A., Irabedra, P., Orlando, D., Rista, L., Haran, G., Viñals, G., ... & Moraña, A. (2003). National prevalence of larval echinococcosis in sheep in slaughtering plants

*Ovis aries* as an indicator in control programmes in Uruguay. *Acta tropica*, 85(2), 281-285.

Cardona, G. A., & Carmena, D. (2013). A review of the global prevalence, molecular epidemiology and economics of cystic echinococcosis in production animals. *Veterinary parasitology*, 192(1), 10-32.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.). New York: Chapman & Hall/CRC Press.

Carlin, B. P., & Louis, T. A. (2008). *Bayesian methods for data analysis*. CRC Press.

Carmena, D., & Cardona, G. A. (2013). Canine echinococcosis: global epidemiology and genotypic diversity. *Acta tropica*, 128(3), 441-460.

Cavagión, L., Alvarez, A., & Larrieu, E. (2002). Diagnóstico histológico del quiste hidatídico ovino y su aplicación en la evaluación de programas de control. In G. Denegri, M.C. Elisondo, & M. Dopschiz. *Situación de la hidatidosis-echinococcosis en la Republica Argentina* (pp.121-131). Buenos Aires: Editorial Martín.

Chuvieco, S. E. (2008). *Teledetección Ambiental: La Observación de la Tierra desde el Espacio*. Barcelona: Ariel.

Colombi, A., Pepe, M., Rampini, A., & Michele, C. De. (2007). Estimation of daily mean air temperature from MODIS LST in Alpine areas. *EARSeL eProceedings*, 6, 563–573.

Coltorti, E. A., & Varela-Diaz, V. M. (1978). Detection of antibodies against *Echinococcus granulosus* arc 5 antigens by double diffusion test. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 72(3), 226-229.

- Coltorti, E. A. (1986). Standardization and evaluation of an enzyme immunoassay as a screening test for the seroepidemiology of human hydatidosis. *The American journal of tropical medicine and hygiene*, 35(5), 1000-1005.
- Congdon, P. (2014). *Applied bayesian modelling* (Vol. 595). John Wiley & Sons.
- Conlon, E. M., & Louis, T. A. (1999). Addressing multiple goals in evaluating region-specific risk using Bayesian methods. In A. B. Lawson, D. Boehning, E. Lasaffree, A. Biggeri, J. F. Viel & R. Bertolline. *Disease mapping and risk assessment for public health* (pp. 31-47). Chichester: fohn Wiley & Sons, Ltd.
- Clayton, D. G., Bernardinelli, L., & Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, 22(6), 1193-1202.
- Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 671-681.
- Clayton, D., & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. *Geographical and environmental epidemiology: methods for small-area studies*, 205-220.
- Clement, E. P. (2014). Small Area Estimation with Application to Disease Mapping. *International Journal of Probability and Statistics*, 3(1), 15-22.
- Craig, P. (1997). Immunodiagnosis of Echinococcus granulosus and a comparison of techniques for diagnoses of canine echinococcosis. In F. Andersen, H. Ouhelli, & M Kachani. *Compendium on echinococcosis in Africa and in a Middle Eastern Countres* (pp. 86-118). Utah: Brigham Young University Print Services.

- Craig, P. S., & Larrieu, E. (2006). Control of cystic echinococcosis/hydatidosis: 1863–2002. *Advances in parasitology*, 61, 443-508.
- Cressie, N. (1992). Smoothing regional maps using empirical Bayes predictors. *Geographical Analysis*, 24(1), 75-95.
- Cringoli, G., Rinaldi, L., Veneziano, V., & Musella, V. (2005). Disease mapping and risk assessment in veterinary parasitology: some case studies. *Parassitologia*, 47(1), 9–25.  
Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16044673>
- Daniel, M., Kolar, J., & Zeman, P. (2004). GIS tools for tick and tick-borne disease occurrence. *Parasitology*, 129(S1), S329-S352.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (Vol. 1). Cambridge university press.
- De La Rocque, S., Michel, V., Plazanet, D., & Pin, R. (2004). Remote sensing and epidemiology: examples of applications for two vector-borne diseases. *Comparative immunology, microbiology and infectious diseases*, 27(5), 331-341.
- Del Carpio, M. A. R. I. O., Moguilansky, S., Costa, M., Panomarenko, H., Bianchi, G., Bendersky, S., ... & Larrieu, E. (2000). Diagnosis of human hydatidosis. Predictive value of a rural ultrasonographic survey in an apparently healthy population. *MEDICINA-BUENOS AIRES-*, 60(4), 466-468.
- Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. London: Arnold.

- Durr, P. A., Tait, N., & Lawson, A. B. (2005). Bayesian hierarchical modelling to enhance the epidemiological value of abattoir surveys for bovine fasciolosis. *Preventive veterinary medicine*, 71(3), 157-172.
- Eckert, J., & Thompson, R. C. A. (1997). Intraspecific variation of *Echinococcus granulosus* and related species with emphasis on their infectivity to humans. *Acta Tropica*, 64(1), 19-34.
- Elliott, P., Wakefield, J. C., Best, N. G., & Briggs, D. J. (2000). *Spatial epidemiology: methods and applications*. Oxford University Press.
- Elliott, P., & Wartenberg, D. (2004). Spatial Epidemiology: Current Approaches and Future Challenges. *Environmental Health Perspectives*, 112(9), 998–1006.
- Flores, F., & Lillo, M. (2010). Simple air temperature estimation method from MODIS satellite images on a regional scale. *Agricultura técnica*, 70(3), 436-445.
- Frider, B., & Larrieu, E. (2010). Treatment of liver hydatidosis: How to treat an asymptomatic carrier?. *World journal of gastroenterology: WJG*, 16(33), 4123.
- Fundamentals of Remote Sensing - Introduction (2015). Natural Resources Canada. Recuperado de <http://www.nrcan.gc.ca/earth-sciences/geomatics/satellite-imagery-air-photos/satellite-imagery-products/educational-resources/9363>.
- Gemmel, M., & Lawson, J. (1995). Epidemiology and control of hydatid disease. In R. Thompson & J. Lymbery. *The biology of Echinococcus granulosus and hydatid disease* (pp. 189-216). London: George Allen and Unwin.

- Gemmell, M.A., Roberts, M.G., Beard & Lawson, J.R., (2001). Epidemiology: Quantitative epidemiology and transmission dynamics with special reference to *Echinococcus granulosus*. In J. Eckert, M.A. Gemmell, F.-X. Meslin, y Z.S. Pawlowski. *WHO/OIE manual on echinococcosis in humans and animals: a public health problem of global concern* (pp. 143 - 154). Paris, France: World Organisation for Animal Health. Retrieved from <http://whqlibdoc.who.int/publications/2001/929044522X.pdf>
- Gelfand, A. E., & Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85(1), 1-11.
- Gelman, A and Rubin, DB (1992) Inference from iterative simulation using multiple sequences, *Statistical Science*, 7, 457-511.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3), 515-534.
- Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15), 2865–2873. doi:10.1002/sim.3107
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). London: Chapman & Hall/CRC.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 721-741.
- GeoDa Center for Geospatial Analysis and Computation (s.f.). Recuperado de <https://geodacenter.asu.edu/projects/opengeoda>.

- Ghosh, M., Natarajan, K., Stroud, T. W. F., & Carlin, B. P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93(441), 273-282.
- Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical science*, 55-76.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, 1, 19.
- Giovanni The Bridge Between Data and Science (s.f). Recuperado de <http://giovanni.gsfc.nasa.gov/giovanni/>.
- Godagnone, R., & Bran, D. (2009). *Inventario integrado de los recursos naturales de la Provincia de Río Negro: geología, hidrogeología, geomorfología, suelo, clima, vegetación y fauna*. Ediciones INTA.[Links].
- Goodchild, M. F. (2000). Communicating geographic information in a digital age. *Annals of the Association of American Geographers*, 90(2), 344-355.
- Guarnera, E. A., Santillan, G., Botinelli, R., & Franco, A. (2000). Canine echinococcosis: an alternative for surveillance epidemiology. *Veterinary parasitology*, 88(1), 131-134.
- Guarnera, E. A. (2009). *Hidatidosis en Argentina: carga de enfermedad*. In *Hidatidosis en Argentina: Carga de enfermedad*. Organización Panamericana de la Salud.
- Han, W. S., Burian, S. J., & Shepherd, J. M. (2011). Assessment of satellite-based rainfall estimates in urban areas in different geographic and climatic regions. *Natural Hazards*, 56(3), 733–747.

- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97-109.
- Hay, S. I., & Lennon, J. J. (1999). Deriving meteorological variables across Africa for the study and control of vector-borne disease: A comparison of remote sensing and spatial interpolation of climate. *Tropical Medicine and International Health*, 4(1), 58–71.
- Herbreteau, V., Salem, G., Souris, M., Hugot, J., & Gonzalez, J. (2005). Sizing up human health through remote sensing: uses and misuses. *Parassitologia*, 47(1), 63.
- Hoffmann, R., Borsboom, G., Saez, M., Dell’Olmo, M. M., Burström, B., Corman, D., ... & Gandarillas, A. (2014). Social differences in avoidable mortality between small areas of 15 European cities: an ecological study. *International journal of health geographics*, 13(1), 1.
- Inskip, H., Beral, V., Fraser, P., & Haskey, J. (1983). Methods for age-adjustment of rates. *Statistics in Medicine*, 2(4), 455-466.
- Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns* (Vol. 70). John Wiley & Sons.
- Jarup, L., Best, N., Toledano, M. B., Wakefield, J., & Elliott, P. (2002). Geographical epidemiology of prostate cancer in Great Britain. *International journal of cancer*, 97(5), 695-699.
- Kalluri, S., Gilruth, P., Rogers, D., & Szczur, M. (2007). Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review. *PLoS Pathog*, 3(10), e116.

- Kelsall, J., & Wakefield, J. (2002). Modeling Spatial Variation in Disease Risk: A Geostatistical Approach. *Journal of the American Statistical Association*, 97(459), 692–701.
- Kéry, M. (2010). *Introduction to WinBUGS for ecologists: Bayesian approach to regression, ANOVA, mixed models and related analyses*. Academic Press.
- Kamenetzky, L., Gutierrez, A. M., Canova, S. G., Haag, K. L., Guarnera, E. A., Parra, A., ... & Rosenzvit, M. C. (2002). Several strains of *Echinococcus granulosus* infect livestock and humans in Argentina. *Infection, Genetics and Evolution*, 2(2), 129-136.
- Kistemann, T., Dangendorf, F., & Schweikart, J. (2002). New perspectives on the use of Geographical Information Systems (GIS) in environmental health sciences. *International journal of hygiene and environmental health*, 205(3), 169-181.
- Larrieu, E., Lester, R., Jáuregui, J., Odriozzola, M., & AGUERO, A. (1986). Epidemiología de la hidatidosis humana en la Provincia de Río Negro. *Acta Gastroent Lat Amer*, 16, 93-108.
- Larrieu, E., Costa, M. T., Cantoni, G., Labanchi, J. L., Bigatti, R., Pérez, A., ... & Romeo, S. (2000). Control program of hydatid disease in the province of Río Negro, Argentina. 1980-1997. *Boletín Chileno de Parasitología*, 55(3/4), 49-53.
- Larrieu, E., Costa, M. T., Cantoni, G., Alvarez, R., Cavagion, L., Labanchi, J. L., ... & Cabrera, P. (2001). Ovine *Echinococcus granulosus* transmission dynamics in the province of Río Negro, Argentina, 1980–1999. *Veterinary parasitology*, 98(4), 263-272.

- Larrieu, E., Del Carpio, M., Costa, M., & Yadon, Z. (2002). Risks factors for hydatidosis in children of Rio Negro Province. A study of cases and control. *Ann. Trop. Med. Parasitol*, 96, 43-52.
- Larrieu, E., Del Carpio, M., Salvitti, J. C., Mercapide, C., Sustersic, J., Panomarenko, H., ... Odriozola, M. (2004). Ultrasonographic diagnosis and medical treatment of human cystic echinococcosis in asymptomatic school age carriers: 5 years of follow-up. *Acta Tropica*, 91(1), 5-13.
- Larrieu, E., & Zanini, F. (2012). Critical analysis of cystic echinococcosis control programs and praziquantel use in South America, 1974-2010. *Revista Panamericana de Salud Pública*, 31(1), 81-87.
- Lawson, J. R., & Gemmell, M. A. (1983). Hydatidosis and cysticercosis: the dynamics of transmission. *Adv Parasitol*, 22, 261-308.
- Lawson, A. B. (1989). Contribution to the 'cancer near nuclear installations' meeting. *J. Roy. Statist. Soc. Series A*, 152, 374-375.
- Lawson, A. B. (1994). Using spatial Gaussian priors to model heterogeneity in environmental epidemiology. *The Statistician*, 69-76.
- Lawson, A. B., Williams, F. L., & Williams, F. (2001). *An introductory guide to disease mapping*. John Wiley.
- Lawson, A. B., Browne, W. J., & Rodeiro, C. L. V. (2003). *Disease mapping with WinBUGS and MLwiN* (Vol. 11). John Wiley & Sons.

- Lawson, A. B. (2009). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. CRC press.
- Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. CRC press.
- Leonard, T., & Hsu, J. S. (1999). *Bayesian methods: an analysis for statisticians and interdisciplinary researchers* (Vol. 5). Cambridge University Press.
- Leyland, A. H., & Davies, C. A. (2005). Empirical Bayes methods for disease mapping. *Statistical Methods in Medical Research*, 14(1), 17-34.
- Machault, V., Vignolles, C., Borchì, F., Vounatsou, P., Pages, F., Briolant, S., ... Rogier, C. (2011). The use of remotely sensed environmental data in the study of malaria. *Geospatial Health*, 5(2), 151–168.
- Machault, V., Vignolles, C., Pagès, F., Gadiaga, L., Tourre, Y. M., Gaye, A., ... Rogier, C. (2012). Risk Mapping of *Anopheles gambiae* s.l. *Densities Using Remotely-Sensed Environmental and Meteorological Data in an Urban Area: Dakar, Senegal*. *PLoS ONE*, 7(11).
- Marshall, R. J. (1991). A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 421-441.
- Mas, J. F. (2011). *Aplicaciones del sensor MODIS para el monitoreo del territorio*. Méjico, ISBN, 978-607.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). CRC press.

Meyer, R. (2014). The BUGS book: A Practical Introduction to Bayesian Analysis. *Journal of Biopharmaceutical Statistics*, 24(3), 706–708.

MODIS Land Surface Temperature and Emissivity (s.f). Recuperado de <http://modis.gsfc.nasa.gov/data/dataproduct/mod11.php>.

MOD13A3 (s.f). Recuperado de [https://lpdaac.usgs.gov/dataset\\_discovery/modis/modis\\_products\\_table/mod13a3](https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod13a3).

Mohebbi, M., Wolfe, R., Jolley, D., Forbes, A. B., Mahmoodi, M., & Burton, R. C. (2011). The spatial distribution of esophageal and gastric cancer in Caspian region of Iran: An ecological analysis of diet and socio-economic influences. *International journal of health geographics*, 10(1), 1.

Mollié, A. (1996). Bayesian mapping of disease. *Markov chain Monte Carlo in practice*. London, UK: Chapman & Hall.

MORAN, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1-2), 17–23.

Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3), 451-482.

MRC Biostatistics Unit. DIC: Deviance Information Criterion (1996-2004). Cambridge, UK. Recuperado de <http://users.jyu.fi/~hemipu/itms/DIC%20web%20site%20from%20BUGS%20project.pdf>.

- MRC Biostatistics Unit (2016). Cambridge, UK. Recuperado de <http://www.mrc-bsu.cam.ac.uk/software/bugs/>.
- MRTWeb (s.f). Recuperado de [https://lpdaac.usgs.gov/data\\_access/mrtweb](https://lpdaac.usgs.gov/data_access/mrtweb).
- Mullner, R. M., Chung, K., Croke, K. G., & Mensah, E. K. (2004). Introduction: geographic information systems in public health and medicine. *Journal of medical systems*, 28(3), 215-221.
- Palaniyandi, M. (2012). The role of Remote Sensing and GIS for spatial prediction of vector-borne diseases transmission: A systematic review. *Journal of vector borne diseases*, 49(4), 197.
- Park, J.H. (2015). CRAN Task View: Bayesian Inference. Recuperado de <https://cran.r-project.org/web/views/Bayesian.html>.
- Pfeiffer, D. U., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., & Clements, A. C., (2008). *Spatial Analysis in Epidemiology*. New York: Oxford University Press.
- Pérez, A., Costa, M. T., Gustavo, C., Mancini, S., Mercapide, C., Herrero, E., ... & Larrieu, E. (2006). Vigilancia epidemiológica de la equinococosis quística en perros, establecimientos ganaderos y poblaciones humanas en la Provincia de Río Negro. *MEDICINA (Buenos Aires)*, 66(3), 193-200.
- Pickle, L. W. (1996). *Atlas of united states mortality*. National Center for Health Statistics, Centers for Disease Control and Prevention, US Dept. of Health and Human Services.
- Puigpinós-Riera, R., Marí-Dell'Olmo, M., Gotsens, M., Borrell, C., Serral, G., Ascaso, C., ... & Sánchez-Villegas, P. (2011). Cancer mortality inequalities in urban areas: a Bayesian

small area analysis in Spanish cities. *International journal of health geographics*, 10(6), 1-10.

Pullan, R. L., Gething, P. W., Smith, J. L., Mwandawiro, C. S., Sturrock, H. J. W., Gitonga, C. W., ... Brooker, S. (2011). Spatial modelling of soil-transmitted helminth infections in Kenya: a disease control planning tool. *PLoS Neglected Tropical Diseases*, 5(2), e958.

Pullan, R. L., Sturrock, H. J. W., Soares Magalhães, R. J., Clements, A. C. a, & Brooker, S. J. (2012). Spatial parasite ecology and epidemiology: a review of methods and applications. *Parasitology*, 139(14), 1870–87.

Rao, J. N. K. (2003). *Small area estimation*. Wiley Inter-science.

Reisen, W. K. (2010). Landscape epidemiology of vector-borne diseases. *Annual Review of Entomology*, 55, 461–83.

Richardson, S., Monfort, C., Green, M., Draper, G., & Muirhead, C. (1995). Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. *Statistics in medicine*, 14(21-22), 2487-2501.

Richardson, S., Jackson, C., & Best, N. (2005). Bayesian hierarchical models in ecological studies. In Invited Paper for the International Workshop for Statistical Modelling, Sydney.

Rinaldi, L., Musella, V., Biggeri, A., & Cringoli, G. (2006). New insights into the application of geographical information systems and remote sensing in veterinary parasitology. *Geospatial Health*, 1(1), 33-47.

- Robert, C., & Casella, G. (2011). A short history of Markov Chain Monte Carlo: subjective recollections from incomplete data. *Statistical Science*, 102-115.
- Rosenzvit, M. C., Zhang, L. H., Kamenetzky, L., Canova, S. G., Guarnera, E. A., & McManus, D. P. (1999). Genetic variation and epidemiology of *Echinococcus granulosus* in Argentina. *Parasitology*, 118(05), 523-530.
- Sartorius, B. K., Sartorius, K., Chirwa, T. F., & Fonn, S. (2011). Infant mortality in South Africa-distribution, associations and policy implications, 2007: an ecological spatial analysis. *International journal of health geographics*, 10(1), 1.
- Schantz, P. M. (1973). Guía para el empleo del bromhidrato de arecolina en el diagnóstico de la infección por *Echinococcus granulosus* en el perro. *Bol Chil Parasitol.*
- Schantz, P. M., Colli, C., Cruz-Reyes, A., & Prezioso, U. (1976). Sylvatic echinococcosis in Argentina. II. Susceptibility of wild carnivores to *Echinococcus granulosus* (Batsch, 1786) and host-induced morphological variation. *Tropenmedizin und Parasitologie*, 27(1), 70-78.
- Schober, E., Rami, B., Waldhor, T., & Karimian-Teherani, D. (2001). Originalarbeiten-Regional distribution of childhood onset diabetes mellitus in Austria--Analysis of national registration from 1989 to 1999. *Wiener Klinische Wochenschrift*, 113(13-14), 491-495.
- Sharma, R. R., Samarkoon, L., Hazarika, M., & Kafle, T. P. (2007). Applicability of tropical rainfall measuring mission to predict floods on the Bagmati River. *J Hydrol Meteorol*, 4(1), 1-15.

- Snow, J. (1854). The cholera near Golden-square, and at Deptford. *Medical Times and Gazette*, 9, 321-322.
- Spiegelhalter, D. J., Best, N. G., Gilks, W. R., & Inskip, H. (1996). Hepatitis B: a case study in MCMC methods. In W. Gilks, S., Richardson, & D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice* (pp. 21-43). Interdisciplinary Statistics Series.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583-639.
- Stevenson, M. A., Morris, R. S., Lawson, A. B., Wilesmith, J. W., Ryan, J. B. M., & Jackson, R. (2005). Area-level risks for BSE in British cattle before and after the July 1988 meat and bone meal feed ban. *Preventive veterinary medicine*, 69(1), 129-144.
- Tatem, A. J., Goetz, S. J., & Hay, S. I. (2004). Terra and Aqua: New data for epidemiology and public health. *International Journal of Applied Earth Observation and Geoinformation*, 6(1), 33-46.
- Thomas, A., Best, N., Lunn, D., Arnold, R., & Spiegelhalter, D. (2004). GeoBugs user manual. Cambridge: Medical Research Council Biostatistics Unit.
- Thompson, R.C.A., & McManus D.P. (2001). Aetiology: parasites and life-cycles. In J. Eckert, M.A. Gemmell, F.-X. Meslin, y Z.S. Pawlowski. *WHO/OIE manual on echinococcosis in humans and animals: a public health problem of global concern* (pp. 1 - 17). Paris, France: World Organisation for Animal Health. Retrieved from <http://whqlibdoc.who.int/publications/2001/929044522X.pdf>

- Toledano, M. B., Jarup, L., Best, N., Wakefield, J., & Elliott, P. (2001). Spatial variation and temporal trends of testicular cancer in Great Britain. *British Journal of Cancer*, 84(11), 1482.
- Torgerson, P. R. (2003). Economic effects of echinococcosis. *Acta Tropica*, 85(2), 113-118.
- Vancutsem, C., Ceccato, P., Dinku, T., & Connor, S. J. (2010). Evaluation of MODIS land surface temperature data to estimate air temperature in different ecosystems over Africa. *Remote Sensing of Environment*, 114(2), 449-465.
- Villagra, S., & Giraudó, C. (2013). Aspectos sistémicos de la producción ovina en la provincia de Río Negro. *Revista Argentina de Producción Animal*, 30(2), 211-224.
- Wakefield, J. C., Best, N. G., & Waller, L. (2000). Bayesian approaches to disease mapping. *Spatial epidemiology: methods and applications*, 104-127.
- Waller, L. A., Carlin, B. P., Xia, H., & Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92(438), 607-617.
- Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data* (Vol. 368). John Wiley & Sons.
- Waller, L. A. (2005). Bayesian thinking in spatial statistics. *Handbook of Statistics*, 25, 589-622.
- Walter, S., Martin Taylor, S., & Marrett, L. (1999). An analysis of determinants of regional variation in cancer incidence: Ontario, Canada. In A. Lawson, A. Biggeri, D. Böhning,

E. Lesaffre, J.F. Viel & R. Bertollini. *Disease mapping and risk assessment for public health* (pp. 365-381). John Wiley & Sons.

Ward, M. P., & Carpenter, T. E. (2000). Techniques for analysis of disease clustering in space and in time in veterinary epidemiology. *Preventive Veterinary Medicine*, 45(3), 257-284.

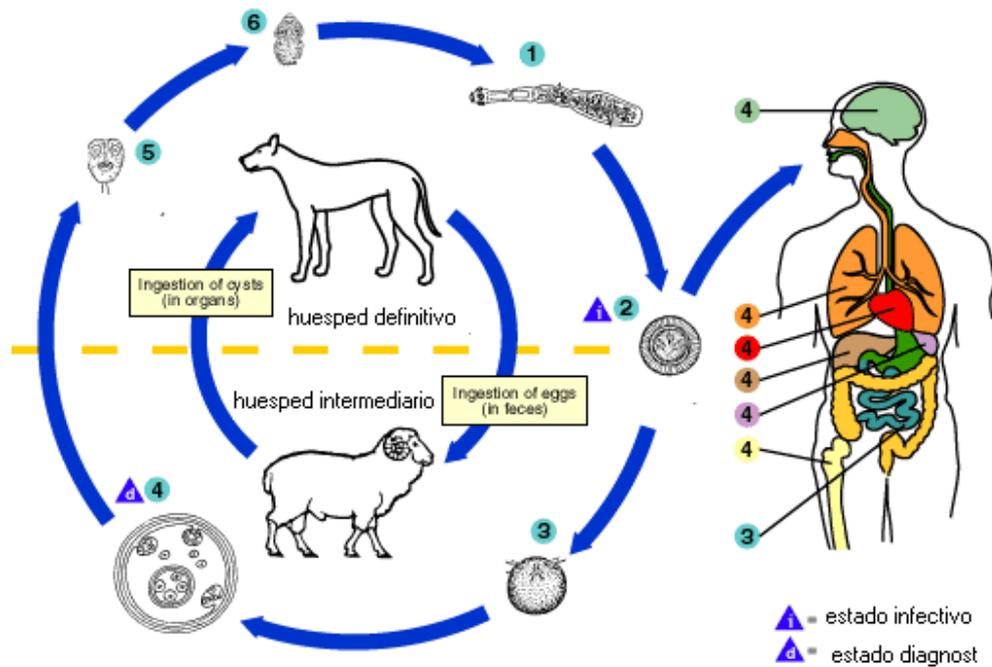
Willgert, K. J., Schroedle, B., & Schwermer, H. (2011). Spatial analysis of bluetongue cases and vaccination of Swiss cattle in 2008 and 2009. *Geospatial health*, 5(2), 227-237.

Xia, H., Carlin, B. P., & Waller, L. A. (1997). Hierarchical models for mapping Ohio lung cancer rates. *Environmetrics*, 8(2), 107-120.

Yildiz, K., & Gurcan, S. (2003). Prevalence of hydatidosis and fertility of hydatid cysts in sheep in Kirikkale, Turkey. *Acta Veterinaria Hungarica*, 51(2), 181-187.

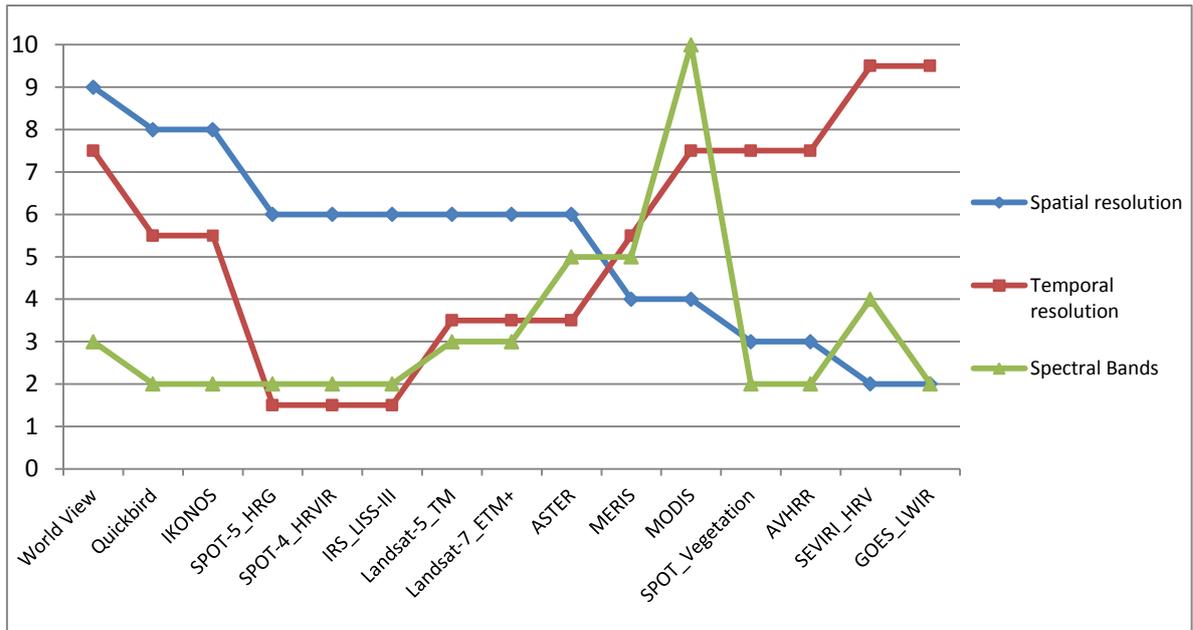
Yu, Q., Scribner, R., Carlin, B., Theall, K., Simonsen, N., Ghosh-Dastidar, B., ... & Mason, K. (2008). Multilevel spatio-temporal dual changepoint models for relating alcohol outlet destruction and changes in neighbourhood rates of assaultive violence. *Geospatial health*, 2(2), 161.

## APÉNDICE A: FIGURAS.



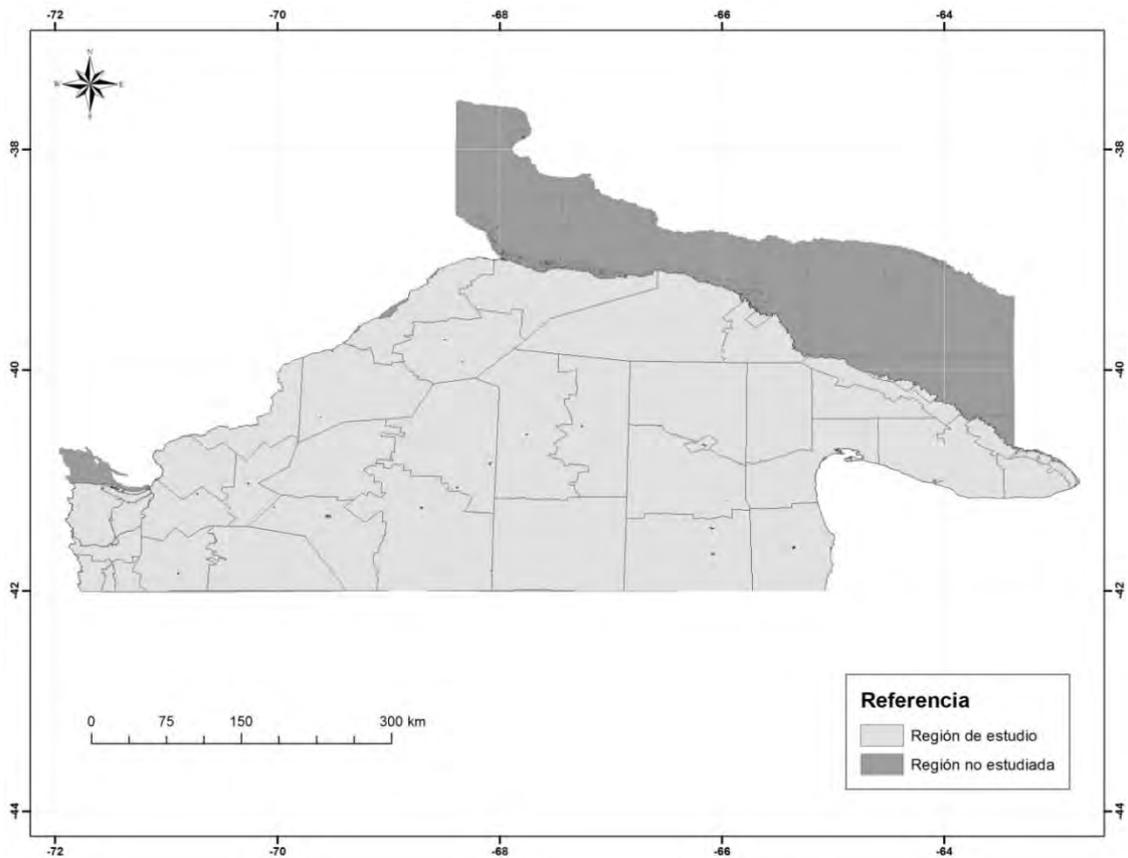
**Figura 1. Esquema del ciclo de transmisión de *Echinococcus granulosus*.**

**Nota Fuente:** Adaptado de Centers for Disease Control and Prevention (C.D.C.). 1: Forma adulta en intestino delgado del perro; 2: oncosfera embrionada en materia fecal; 3 oncosfera penetra en pared intestinal del hospedador intermediario; 4: principales localizaciones del metacestode: hígado y pulmón; 5: protoescólices intraquistos; 6: escólices enganchados en pared intestinal.

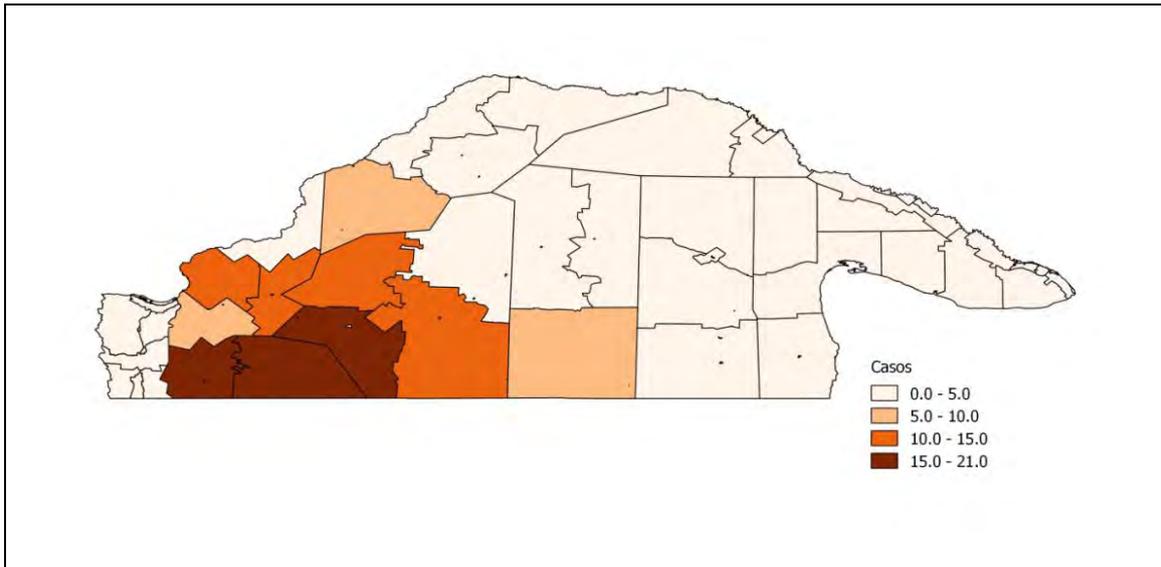


**Figura 2. Muestra la relación entre la resolución espacial, temporal y espectral de sensores satelitales usados o de potencial uso en epidemiología.**

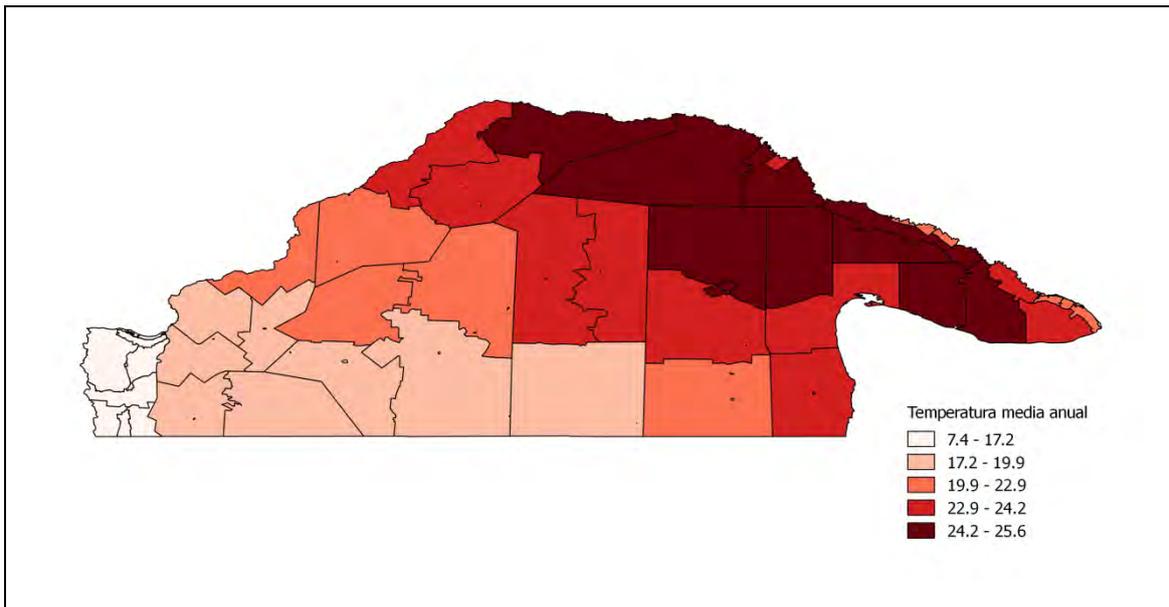
Fuente: Arezo, 2012. Los valores del eje “y” se encuentran estandarizados y ponderados, para mayor detalle ver fuente.



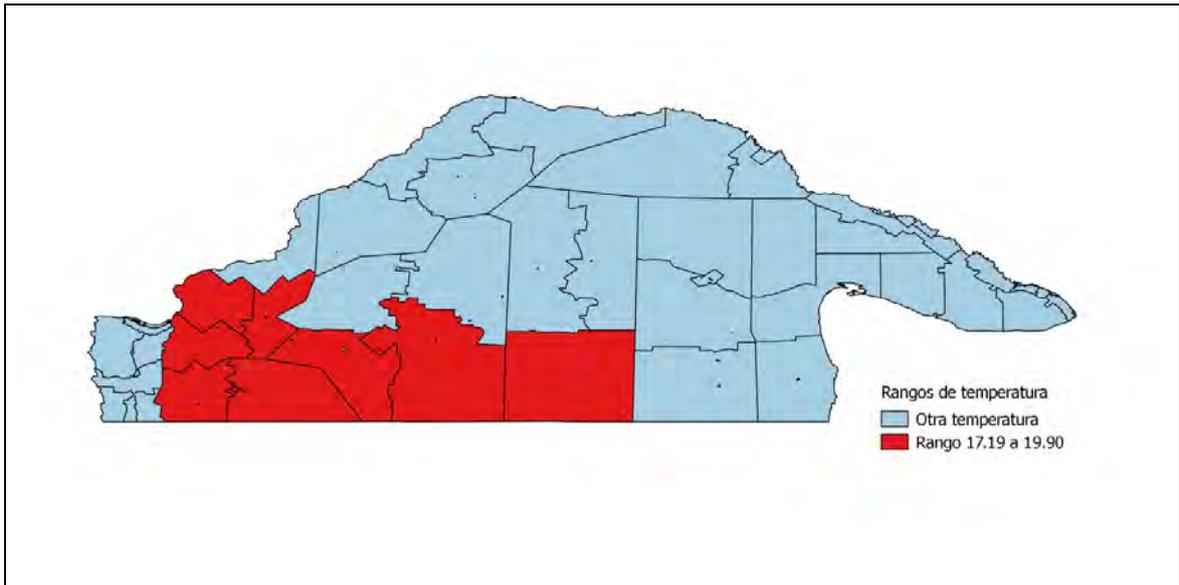
**Figura 3. Región de estudio, provincia de Río Negro.**



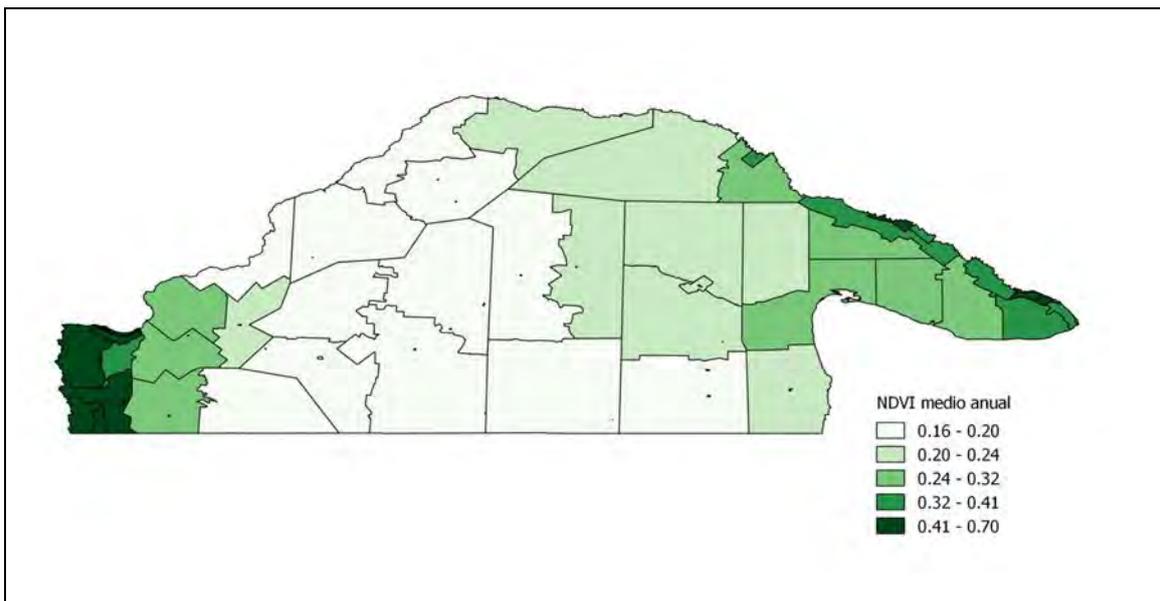
**Figura 4. Mapa de casos de hidatidosis en chicos de 0 a 14 años, provincia de Río Negro.**



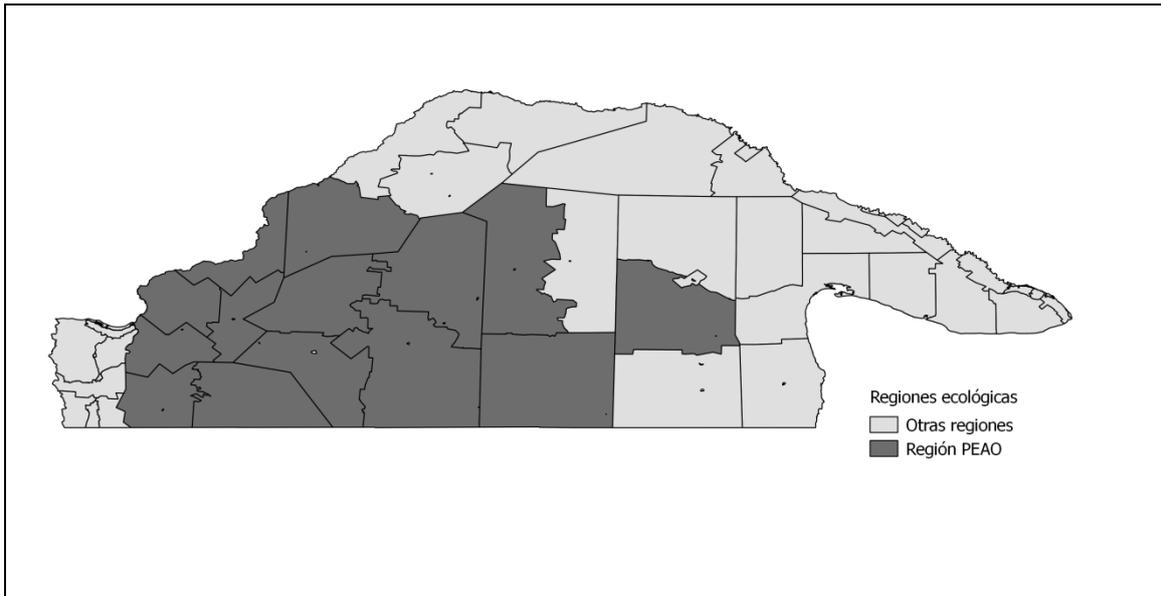
**Figura 5. Mapa de rangos de temperatura media anual (LST) en grados Celsius.**



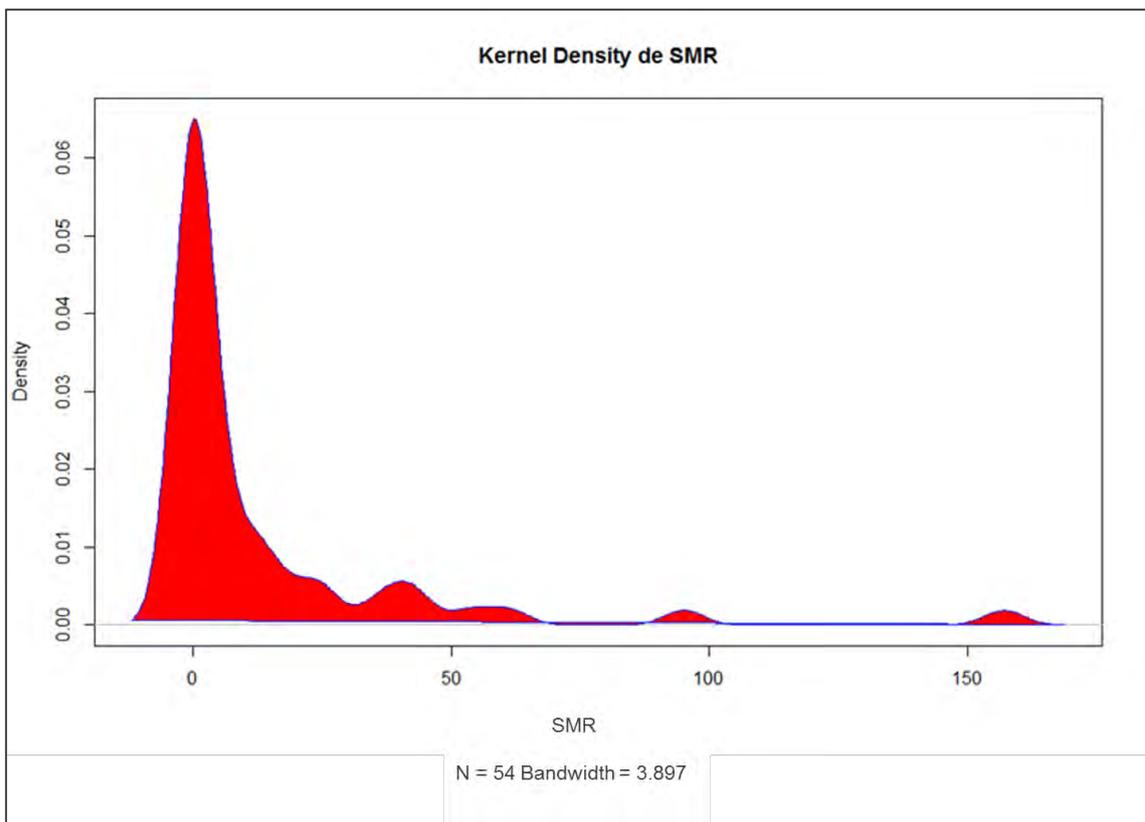
**Figura 6. Mapa de temperatura media anual (LST) para el rango de 17.19 a 19.90 °C y otros rangos.**



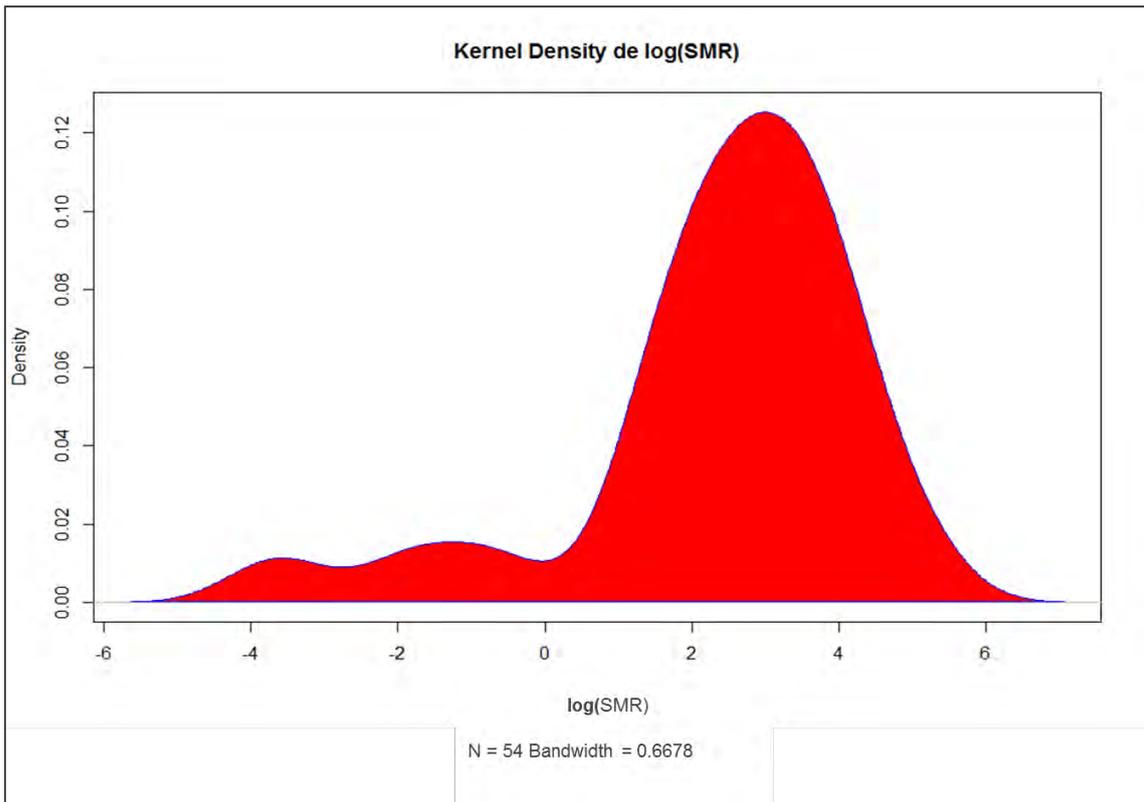
**Figura 7. Mapa de NDVI medio anual.**



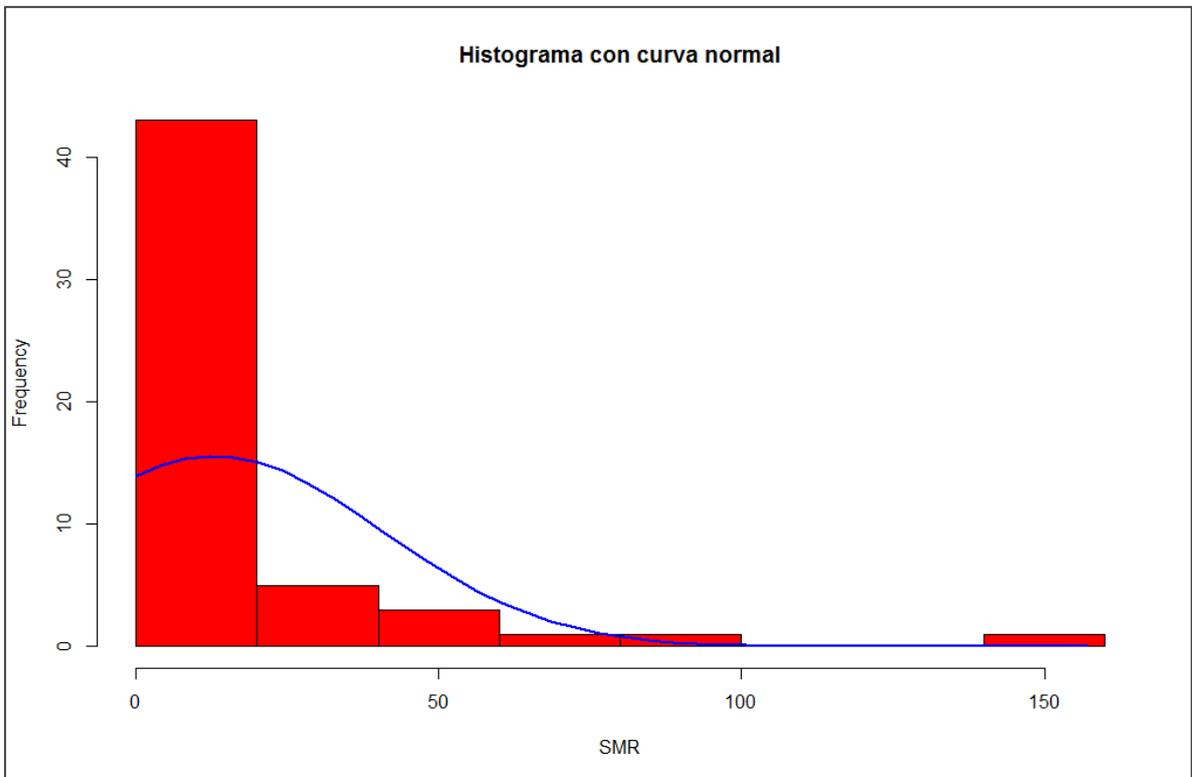
**Figura 8. Mapa de la región ecológica PEAO y otras regiones.**



**Figura 9. Densidad de distribución de la SMR de hidatidosis en chico de 0 a 14 años.**



**Figura 10. Densidad de distribución del logaritmo de la SMR de hidatidosis en chicos de 0 a 14 años.**



**Figura 11. Histograma de SMR con curva normal superpuesta.**

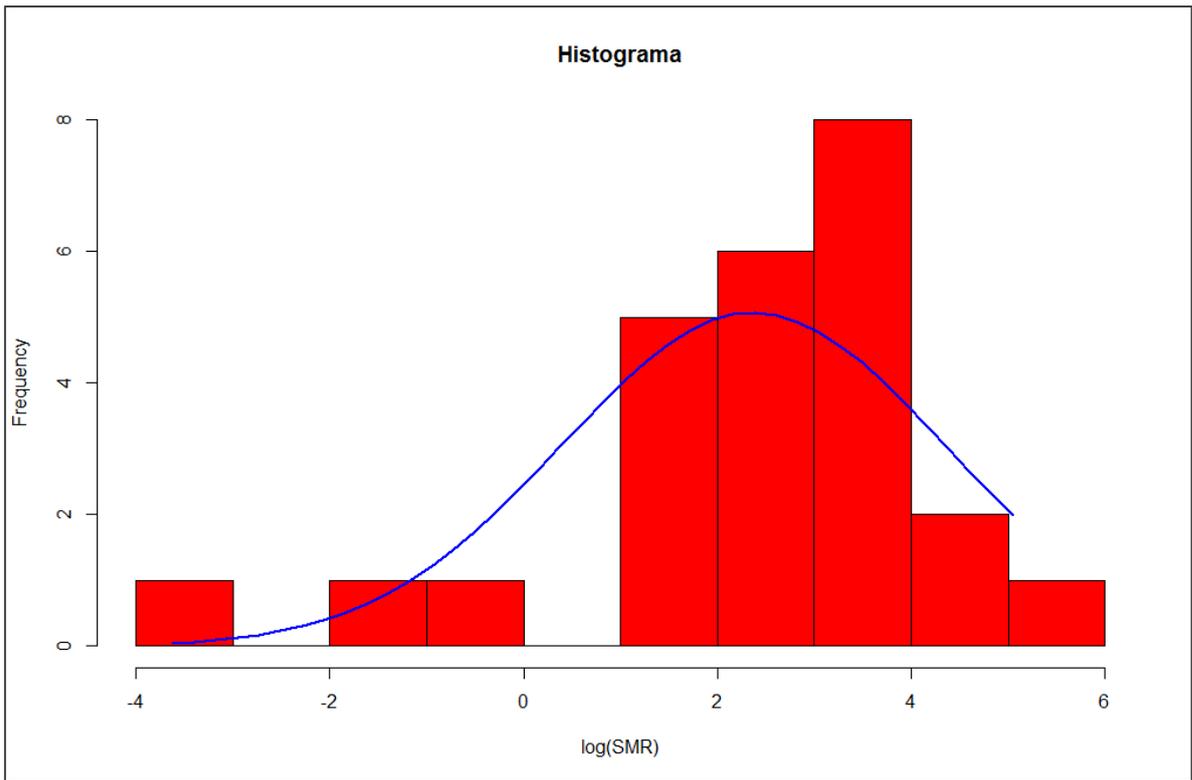


Figura 12. Histograma de log(SMR) con curva normal superpuesta.

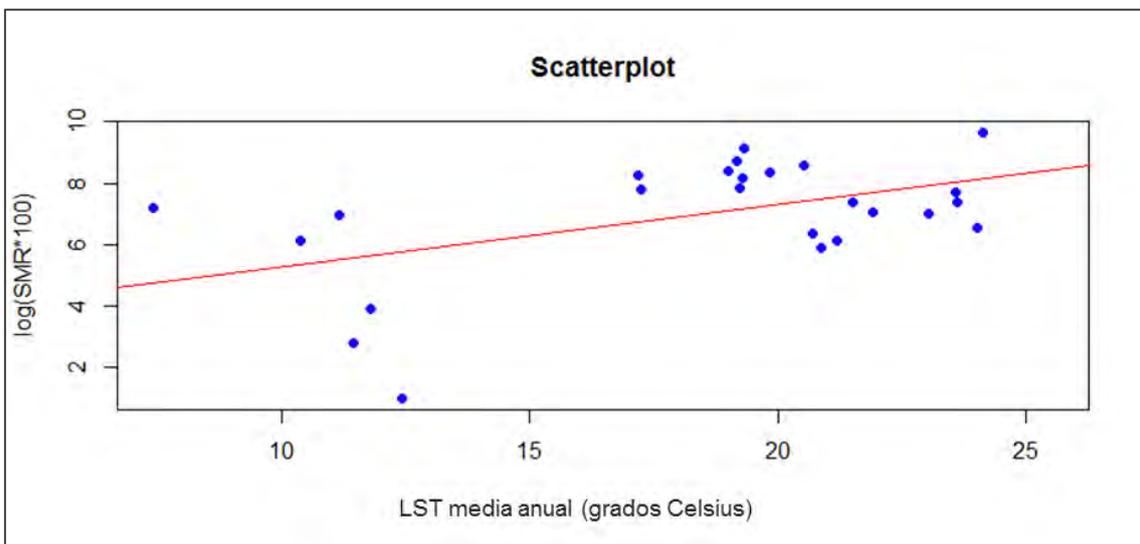


Figura 13. Gráfico del porcentaje del logaritmo de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus temperatura media anual (eje x).

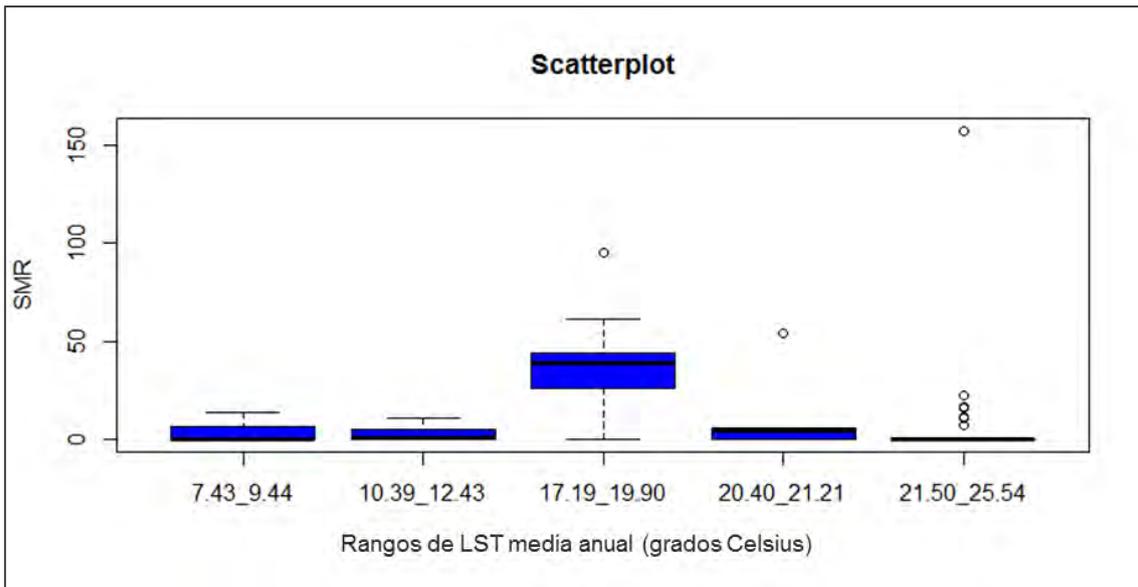


Figura 14. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus rangos de temperatura media anual (eje x).

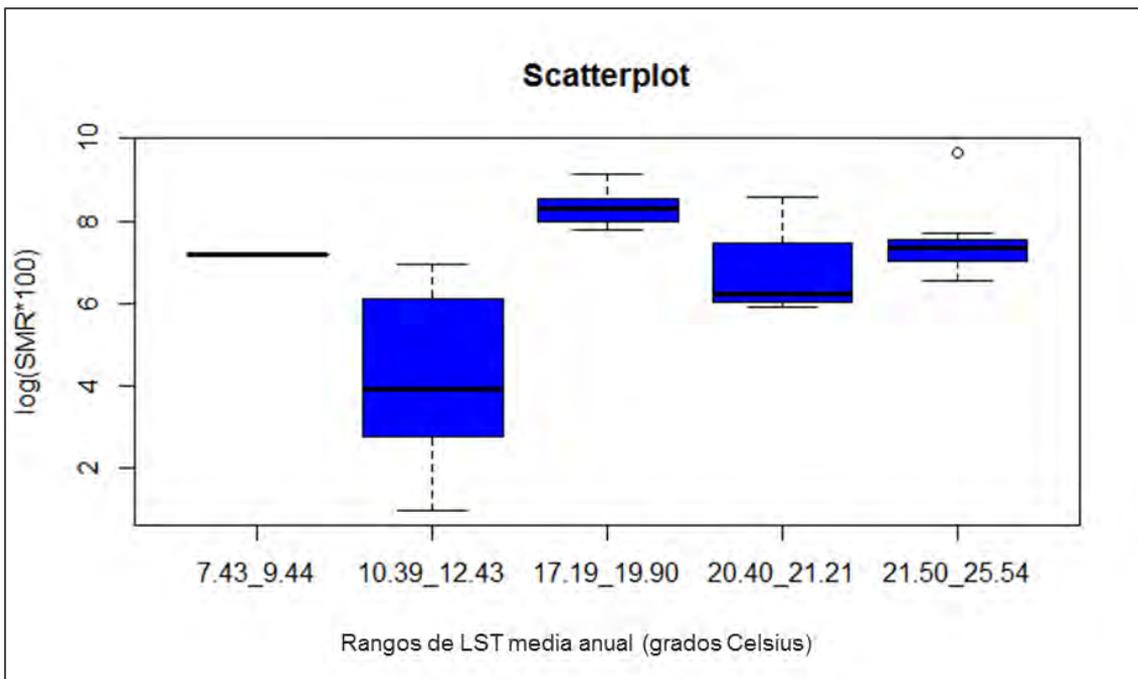


Figura 15. Gráfico del porcentaje del logaritmo de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus rangos de temperatura media anual (eje x).

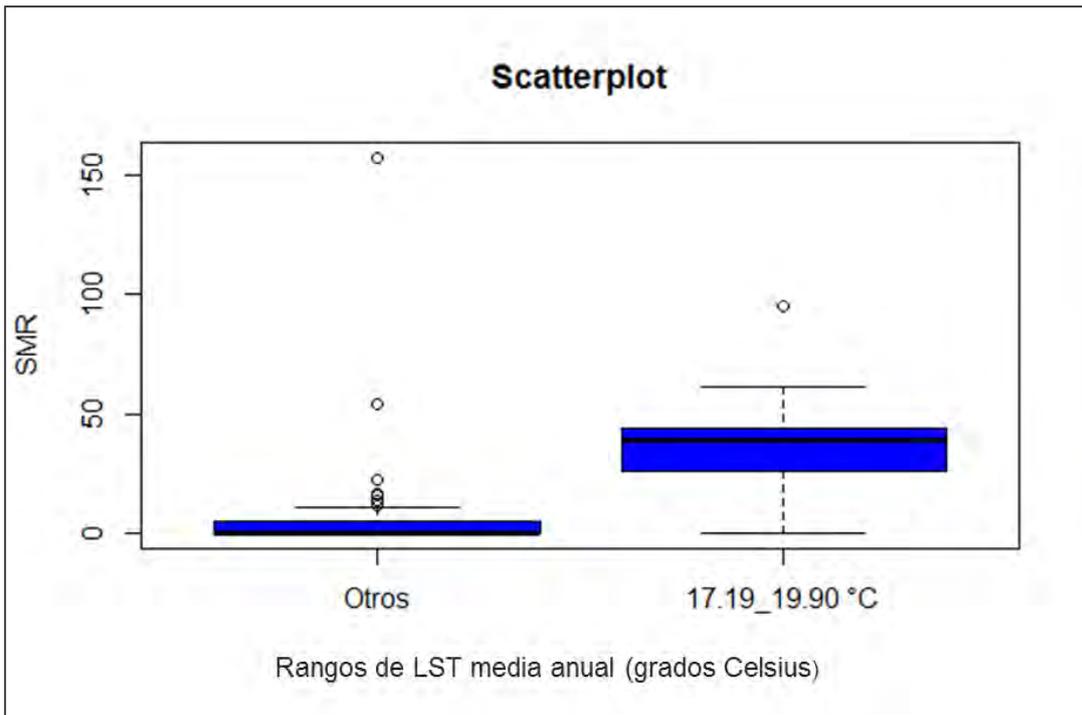


Figura 16. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus rango de temperatura media anual (eje x).

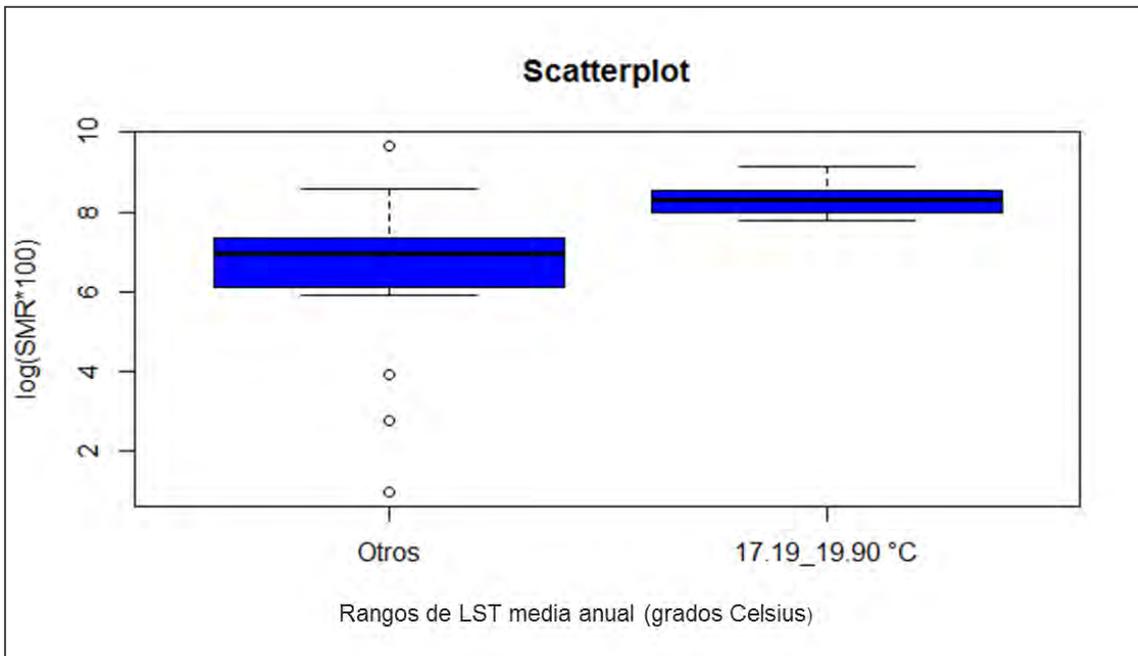


Figura 17. Gráfico del logaritmo del porcentaje de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus rango de temperatura media anual (eje x).

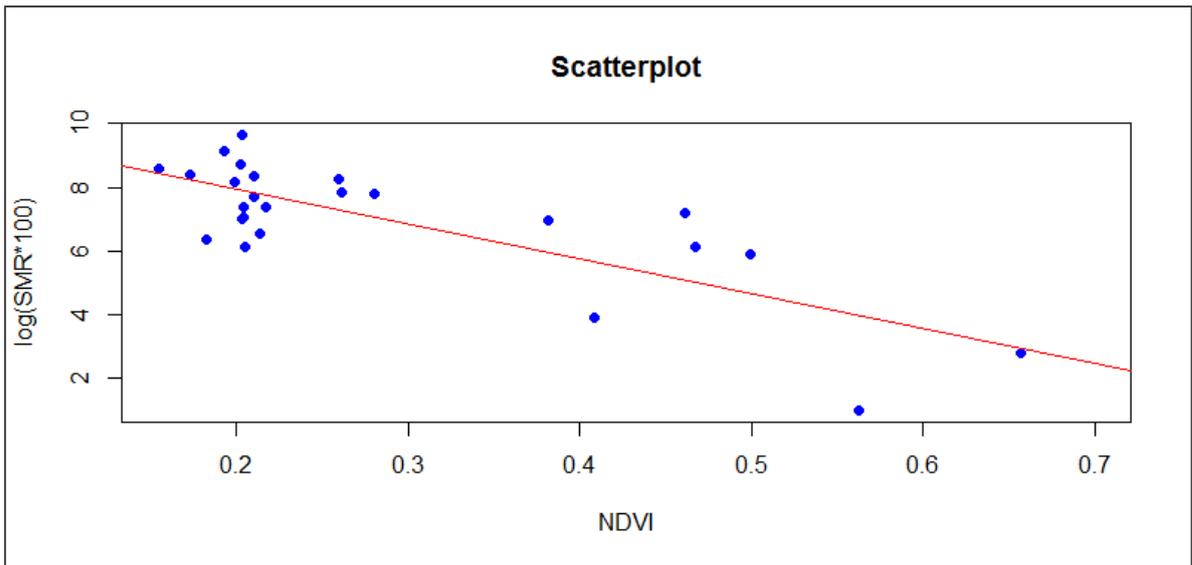


Figura 18. Gráfico del porcentaje del logaritmo de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus NDVI medio anual (eje x).

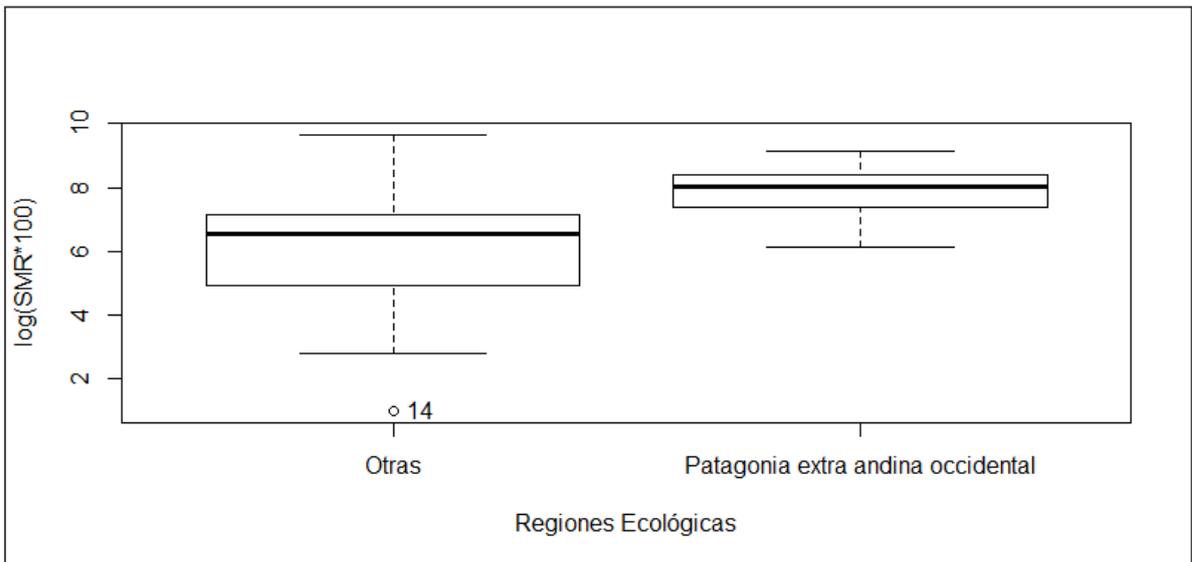


Figura 19. Gráfico del logaritmo del porcentaje de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus regiones ecológicas (eje x).

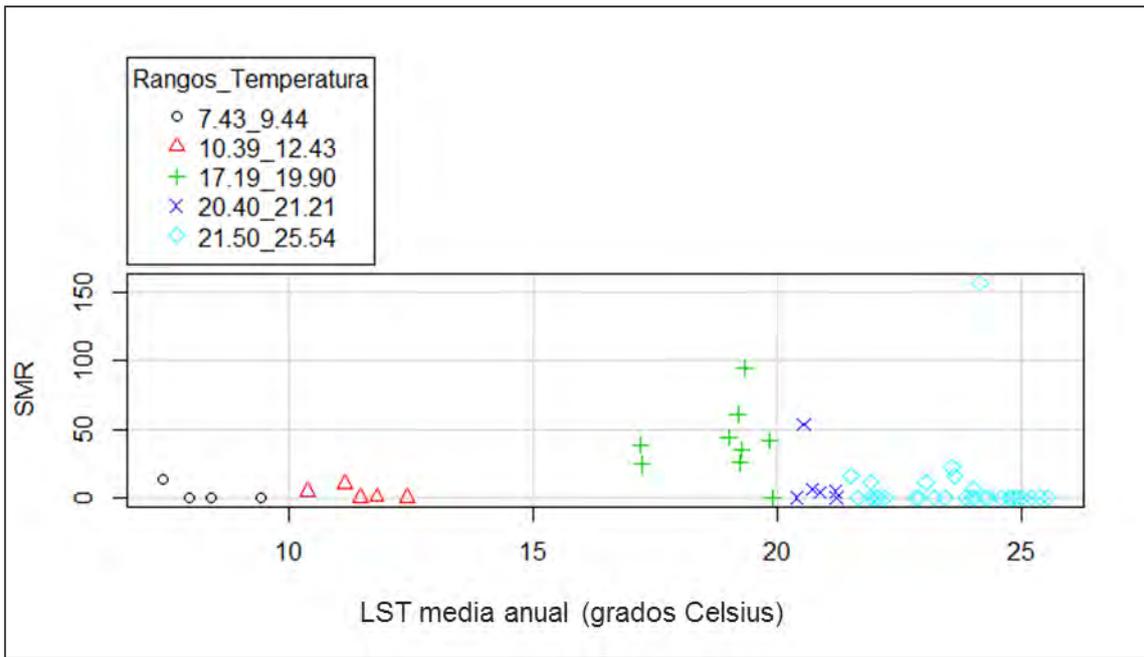


Figura 20. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus temperatura media anual (eje x), distinguida por rangos de temperatura.

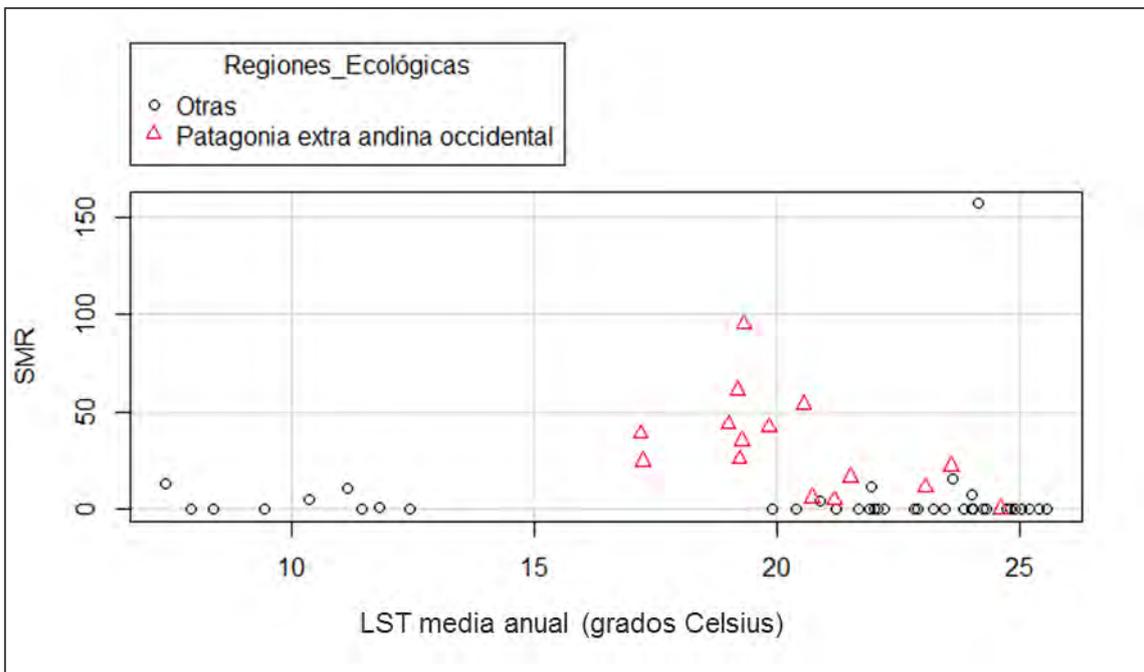


Figura 21. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus temperatura media anual (eje x) y distinción en regiones ecológicas PEAO y otras.

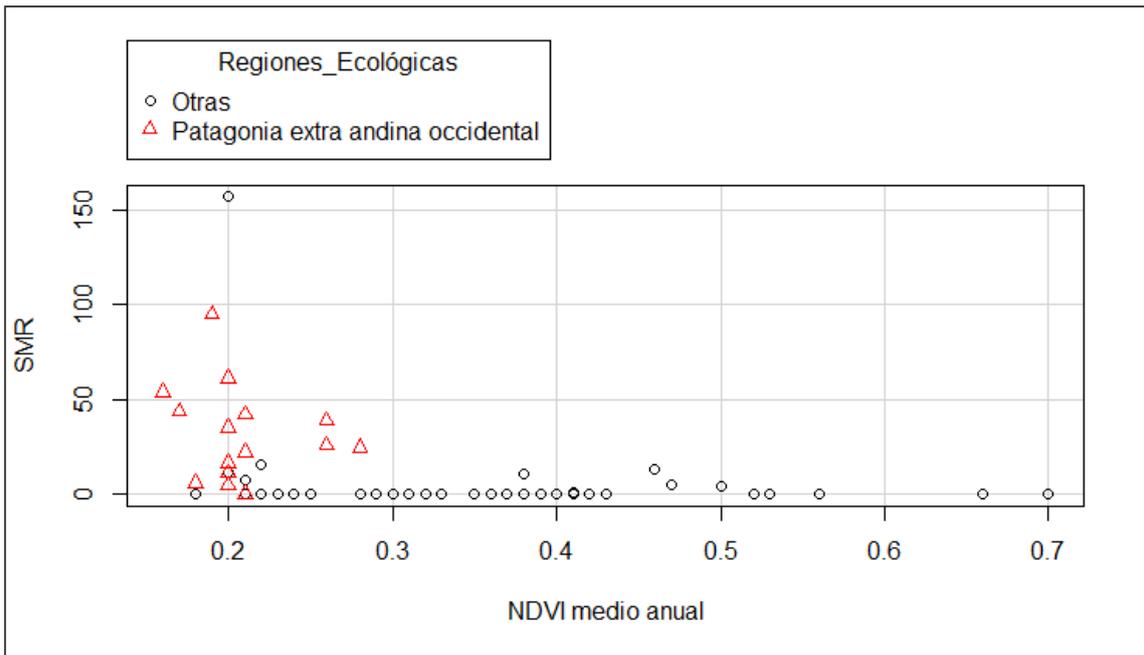


Figura 22. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus NDVI medio anual (eje x) y distinción en regiones ecológicas PEAO y otras.

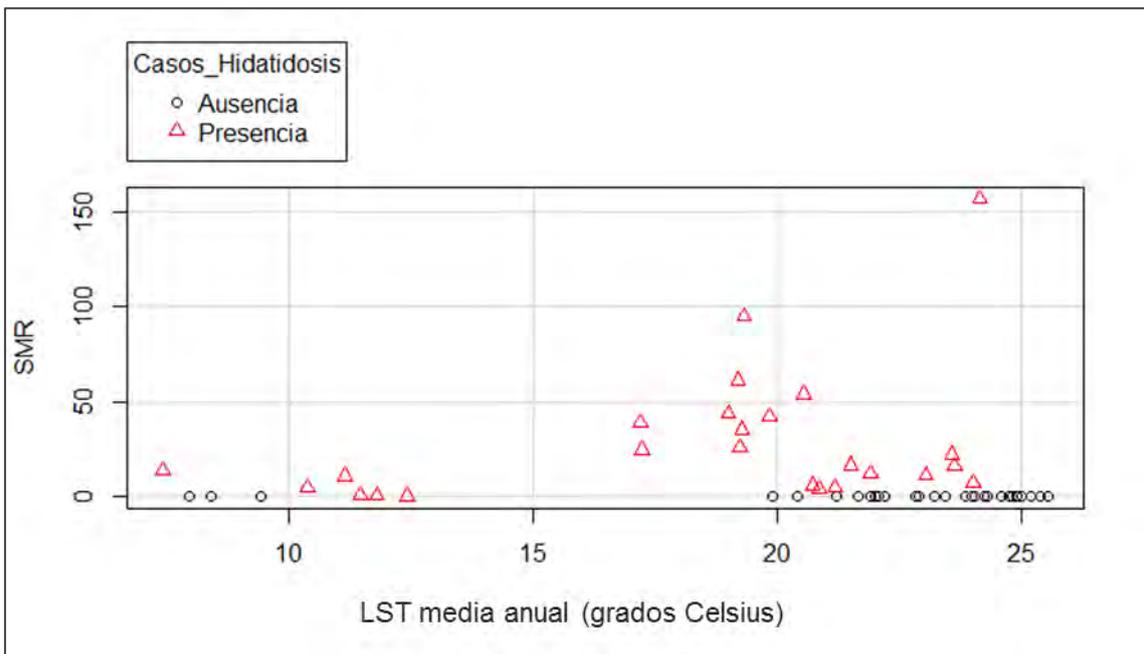
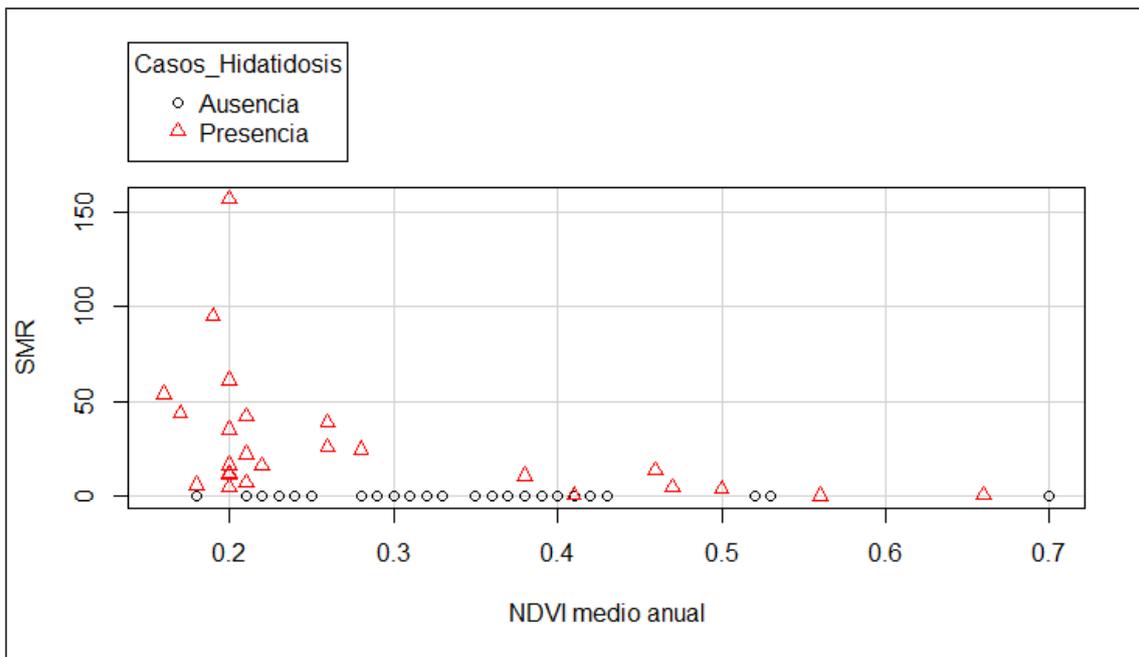
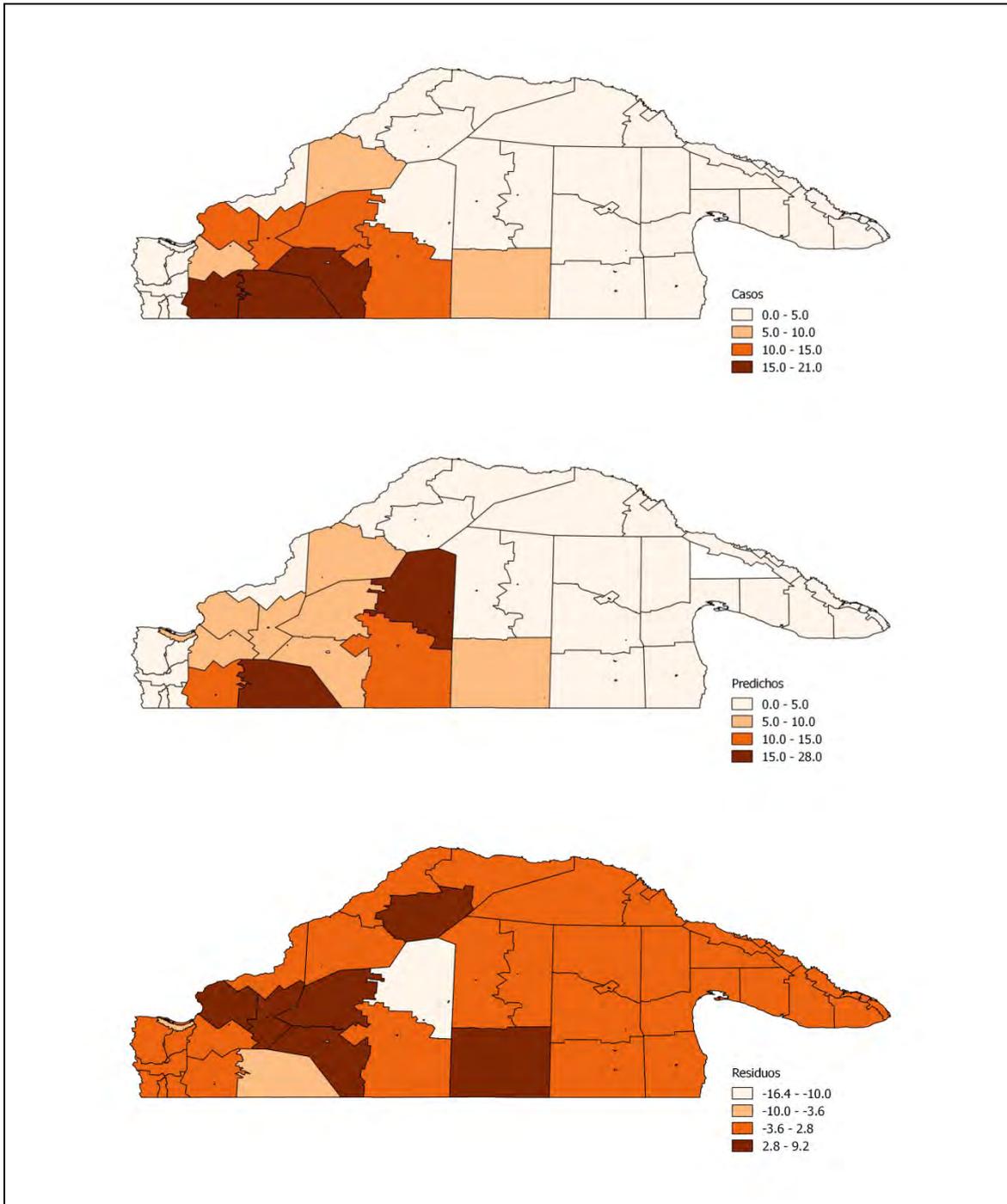


Figura 23. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus temperatura media anual (eje x) y distinción por presencia y ausencia de casos de hidatidosis en chicos de 0 a 14 años.



**Figura 24. Gráfico de la SMR (eje y) de hidatidosis en chicos de 0 a 14 años, versus NDVI medio anual (eje x) y distinción por presencia y ausencia de casos de hidatidosis en chicos de 0 a 14 años.**



**Figura 25.** Se muestra el mapa de casos observados (superior), casos predichos por el modelo (intermedio) y residuos (inferior), obtenidos por la diferencia de los anteriores, para la región de estudio y por radio censal, para el modelo MOD\_1.

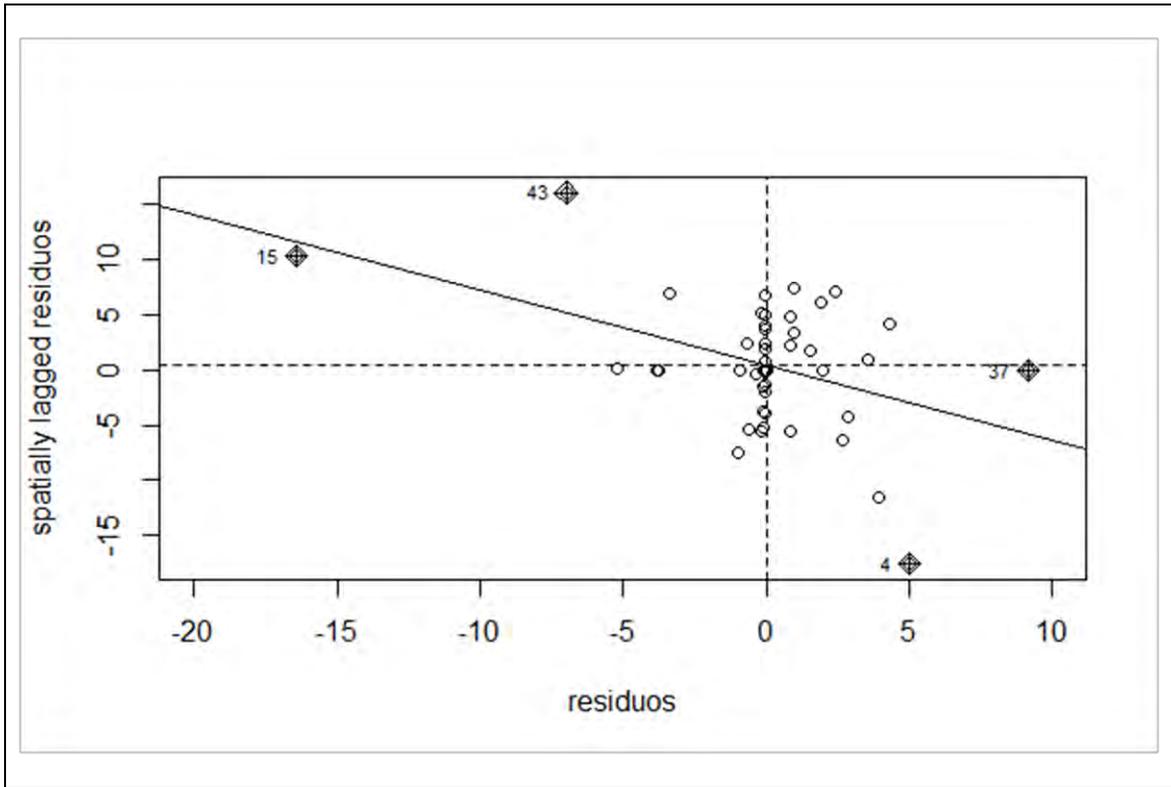


Figura 26. Moran's I scatterplot

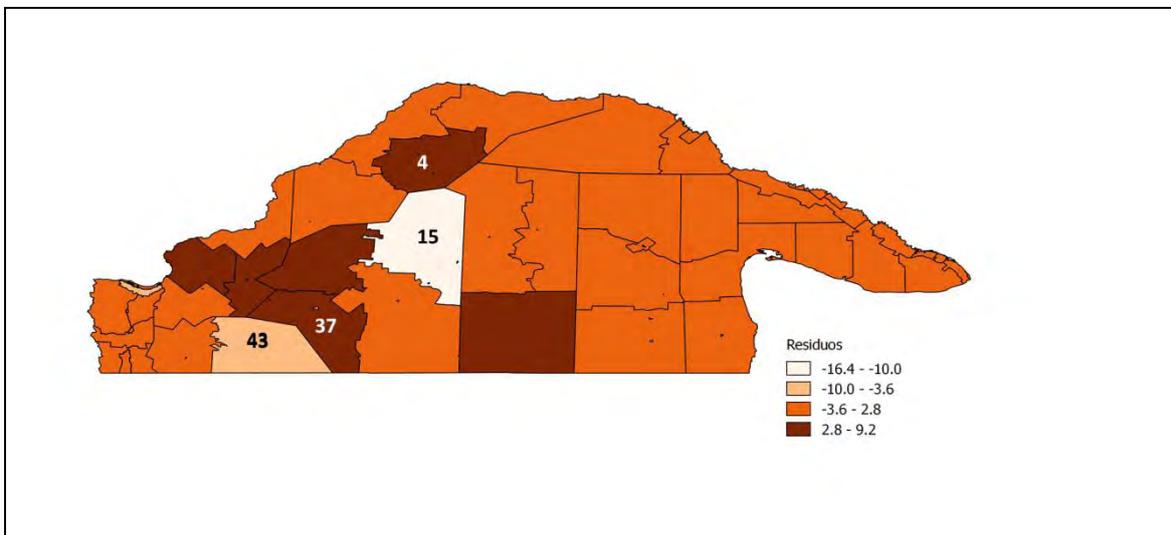


Figura 27. Radios censales Outlier: 4, 15, 37 y 43.

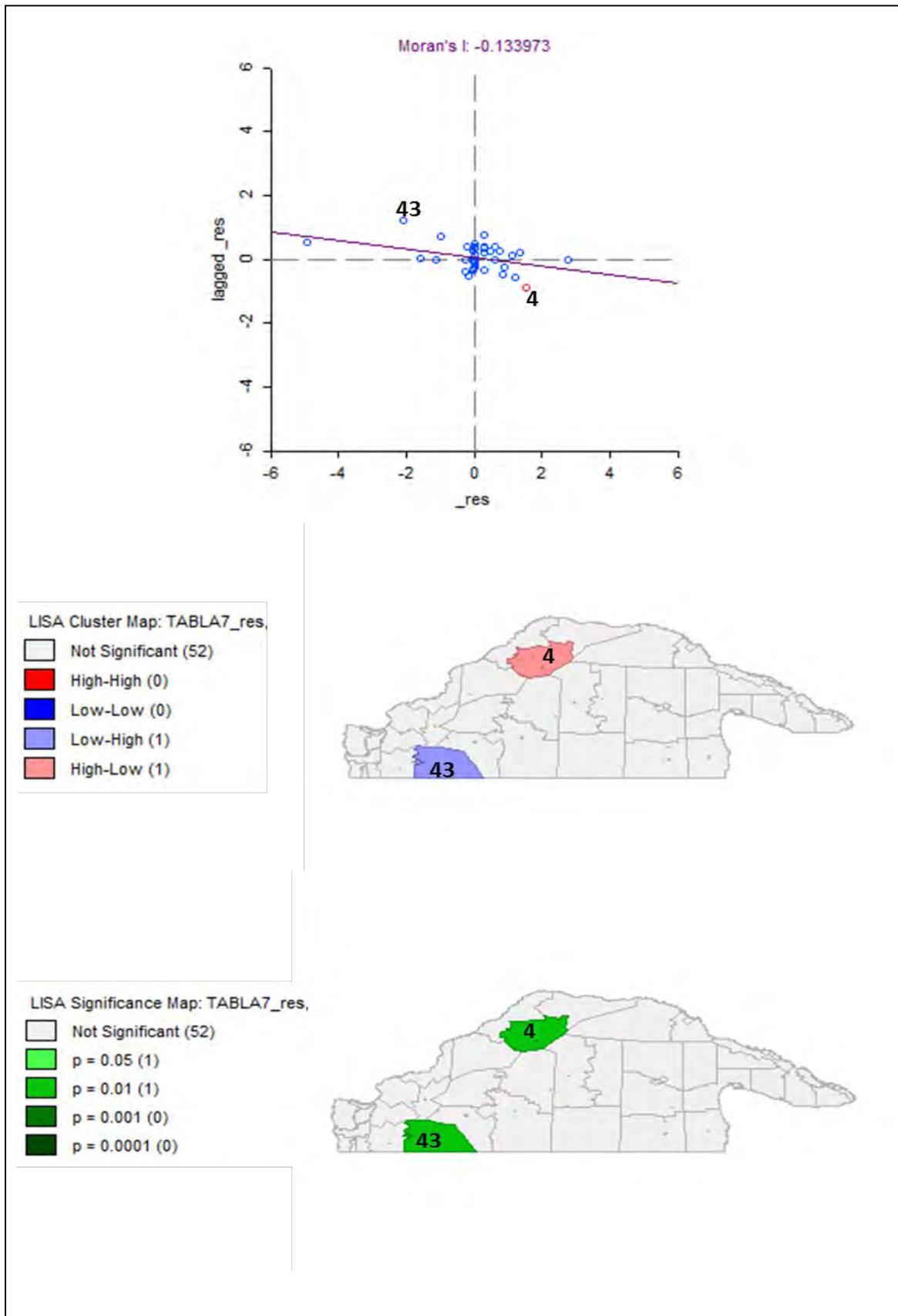
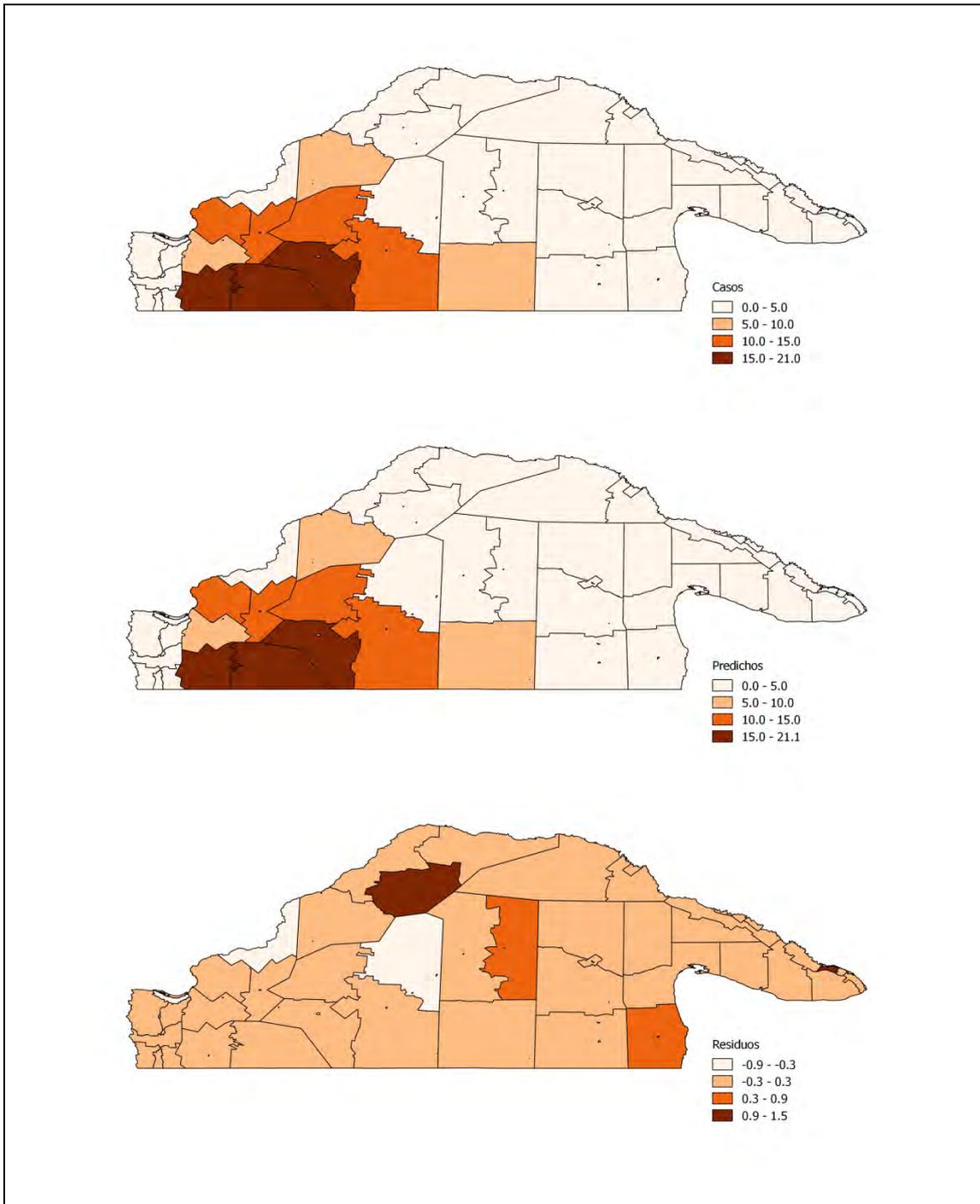


Figura 28. Indicadores locales de asociación espacial para los residuos de MOD\_1.



**Figura 29. Mapas de casos observados (superior), casos predichos por el modelo (intermedio) y residuos (inferior) para la región de estudio y por radio censal, para el modelo MOD\_3.**

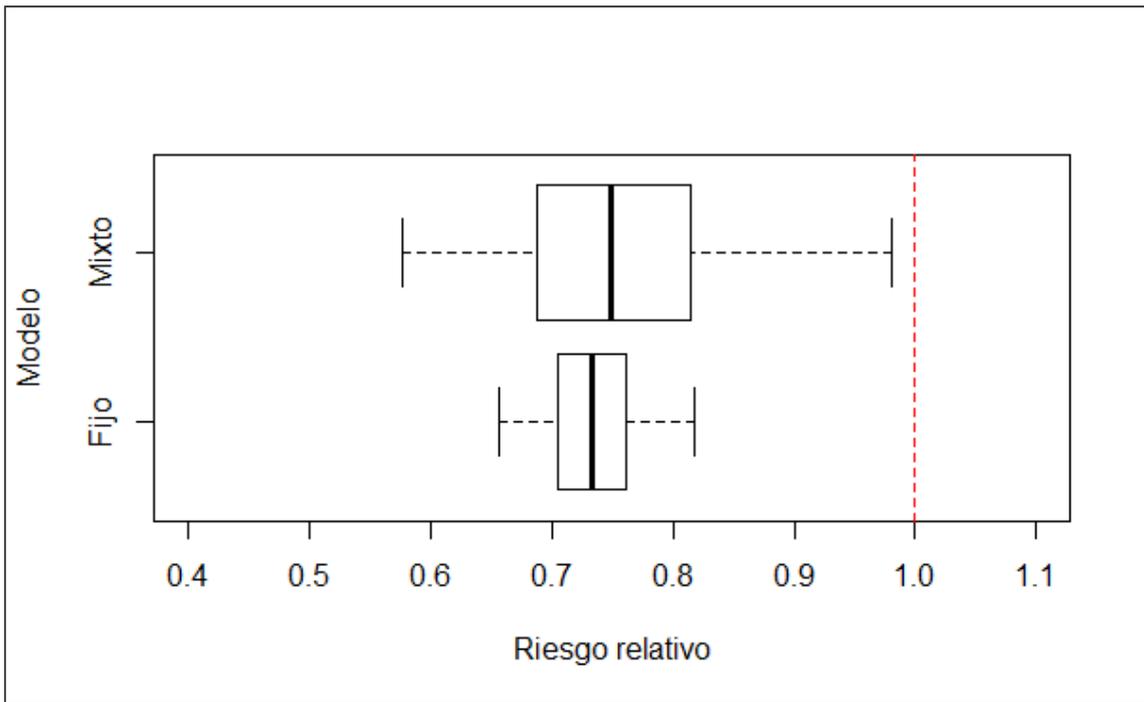


Figura 30. Boxplot mostrando el intervalo de credibilidad del 95% para la estimación del riesgo relativo de los efectos de la temperatura media anual de la superficie terrestre (Grados Celsius).

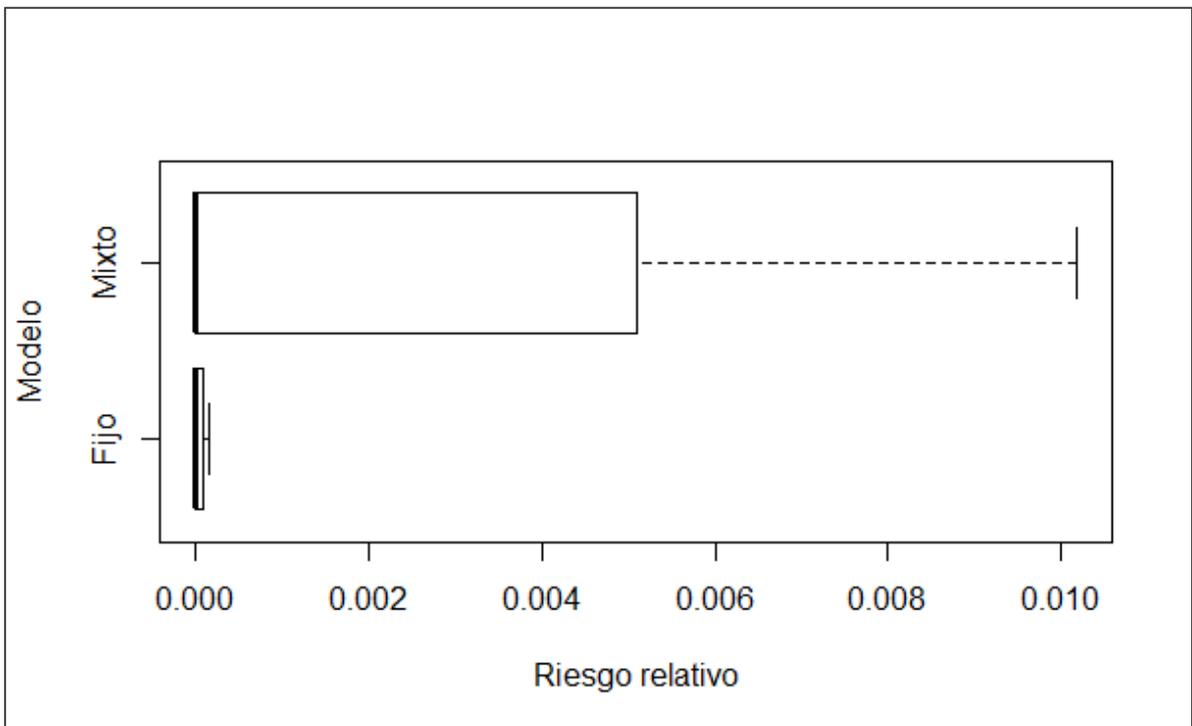
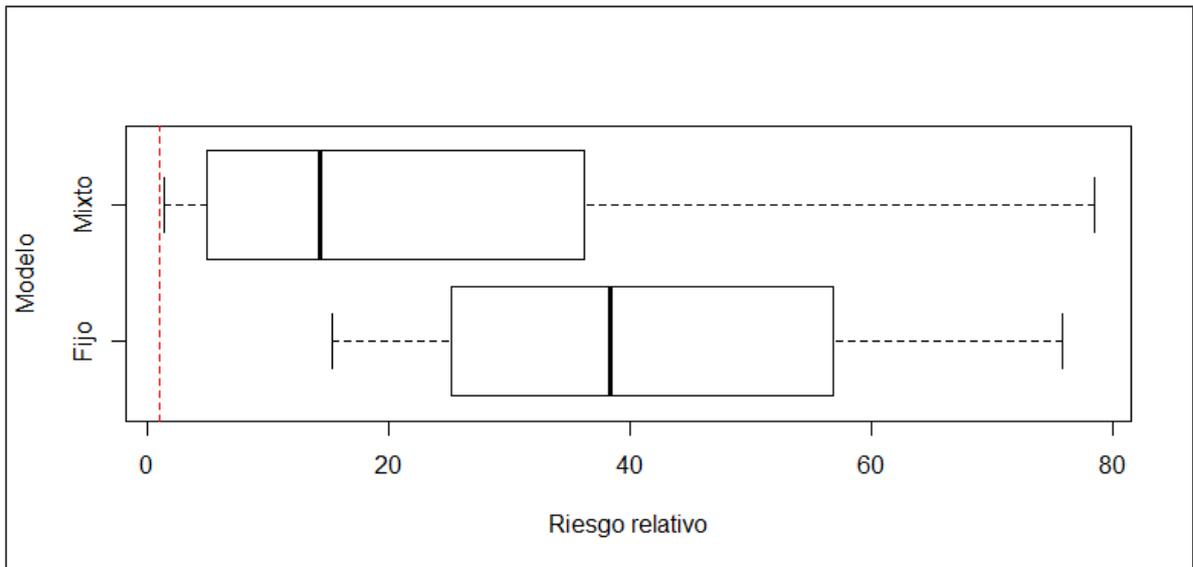
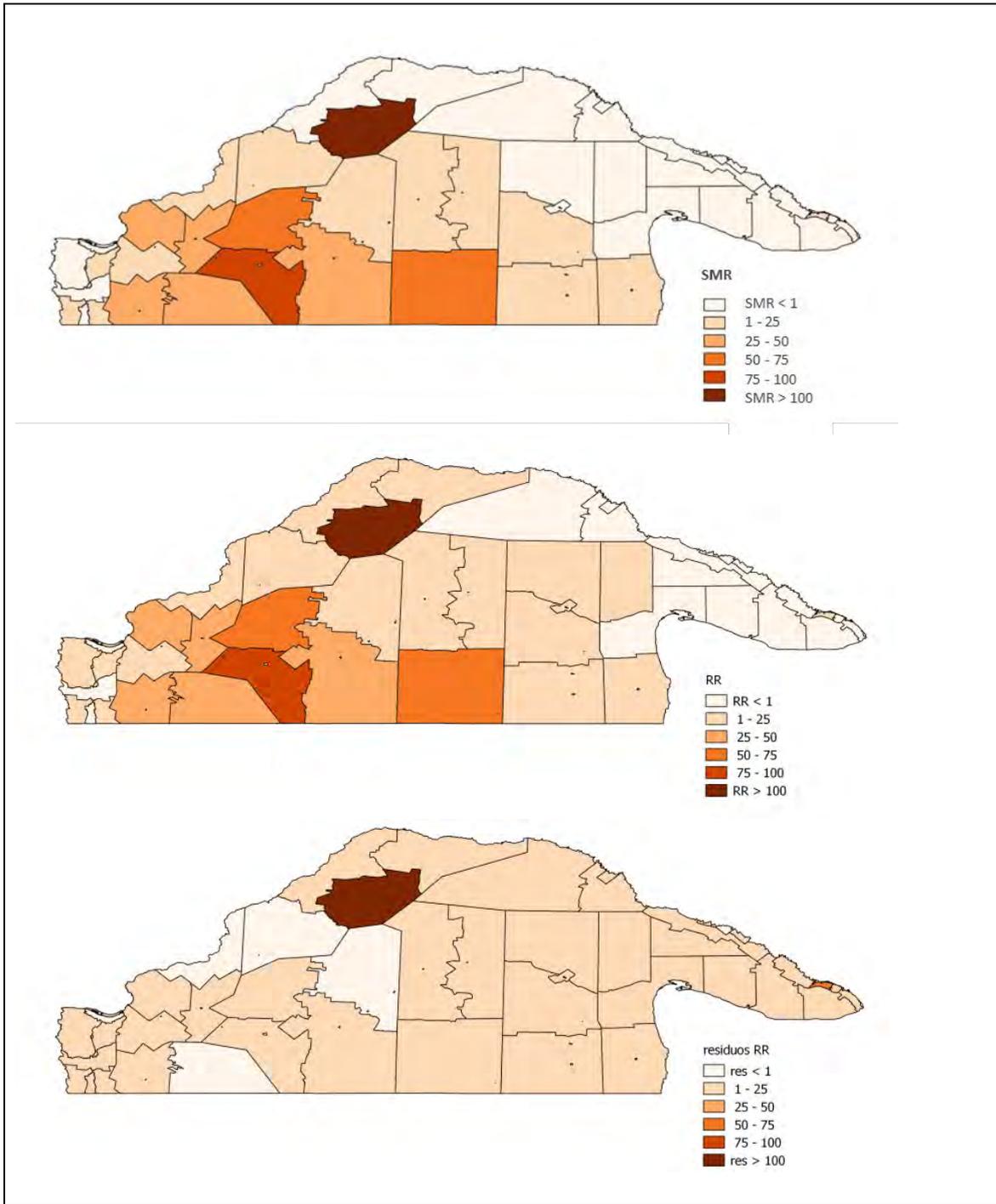


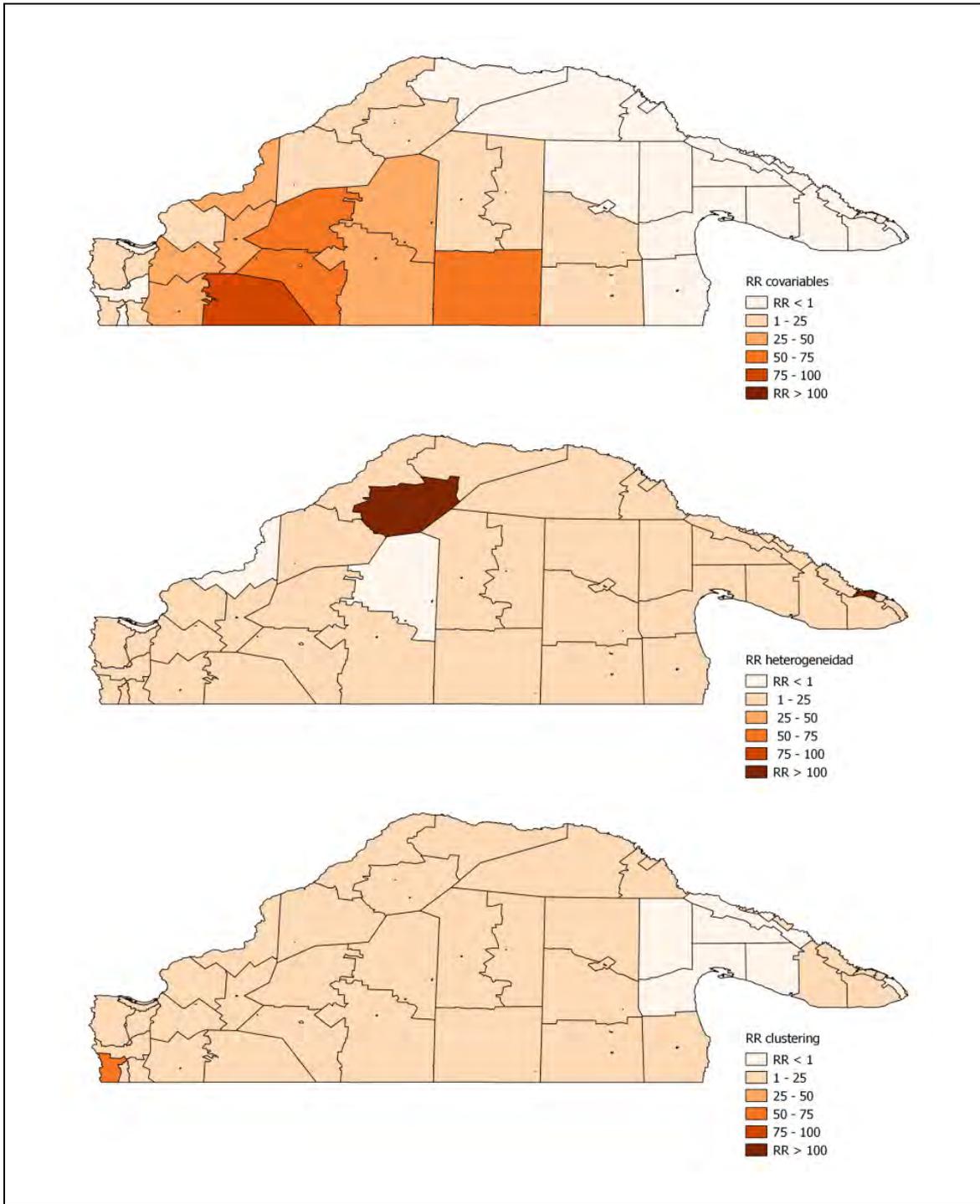
Figura 31. Boxplot mostrando el intervalo de credibilidad del 95% para la estimación del riesgo relativo de los efectos del NDVI.



**Figura 32. Boxplot mostrando el intervalo de credibilidad del 95% para la estimación del riesgo relativo de los efectos de la región ecológica PEAO.**



**Figura 33. SMR de hidatidosis en niños de 0 a 14 años (superior), el RR fijado por el modelo de EM (intermedio) y los residuos del RR, representados por radio censal para la región de estudio (inferior).**



**Figura 34. Componentes de la variación de la morbilidad de Hidatidosis en chicos de 0 a 14 años.**

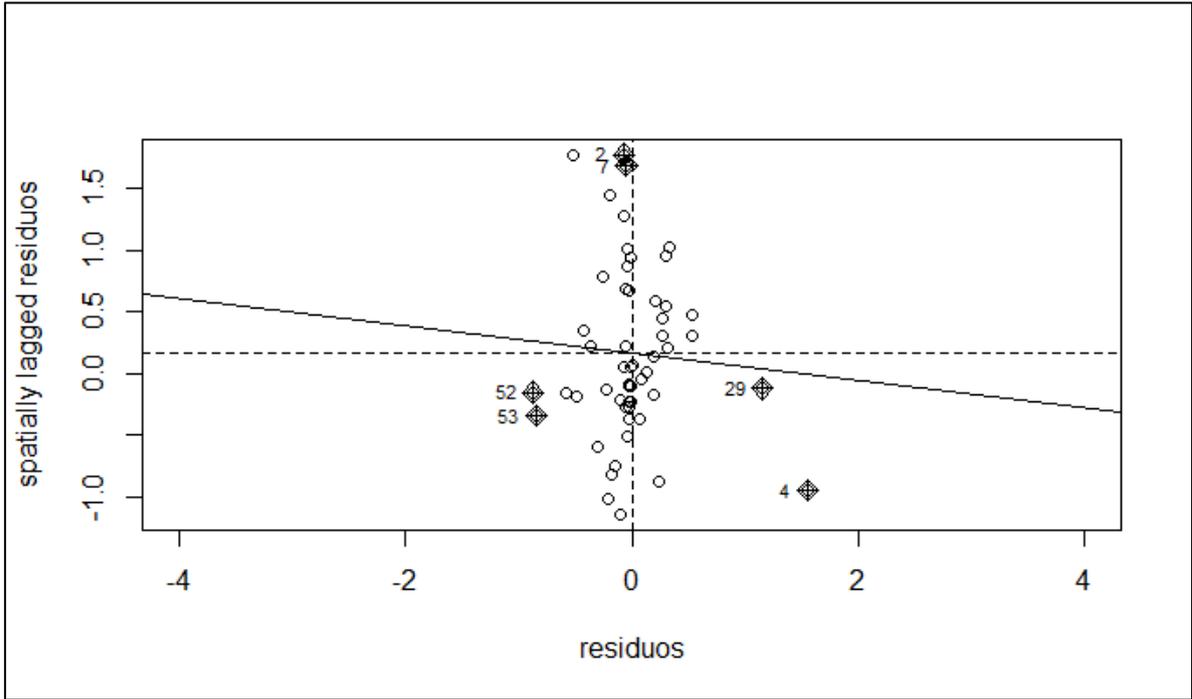


Figura 35. Moran's I scatterplot del modelo de EM con efectos aleatorios incorporados (MOD\_3).

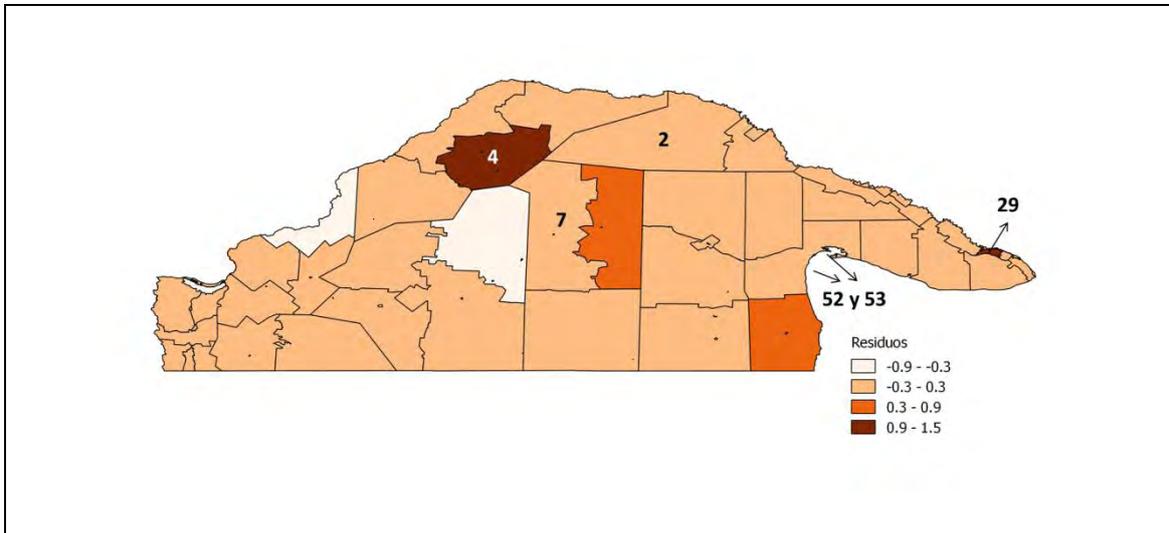


Figura 36. Mapa de residuos del modelo MOD\_3, mostrando los valores atípicos.

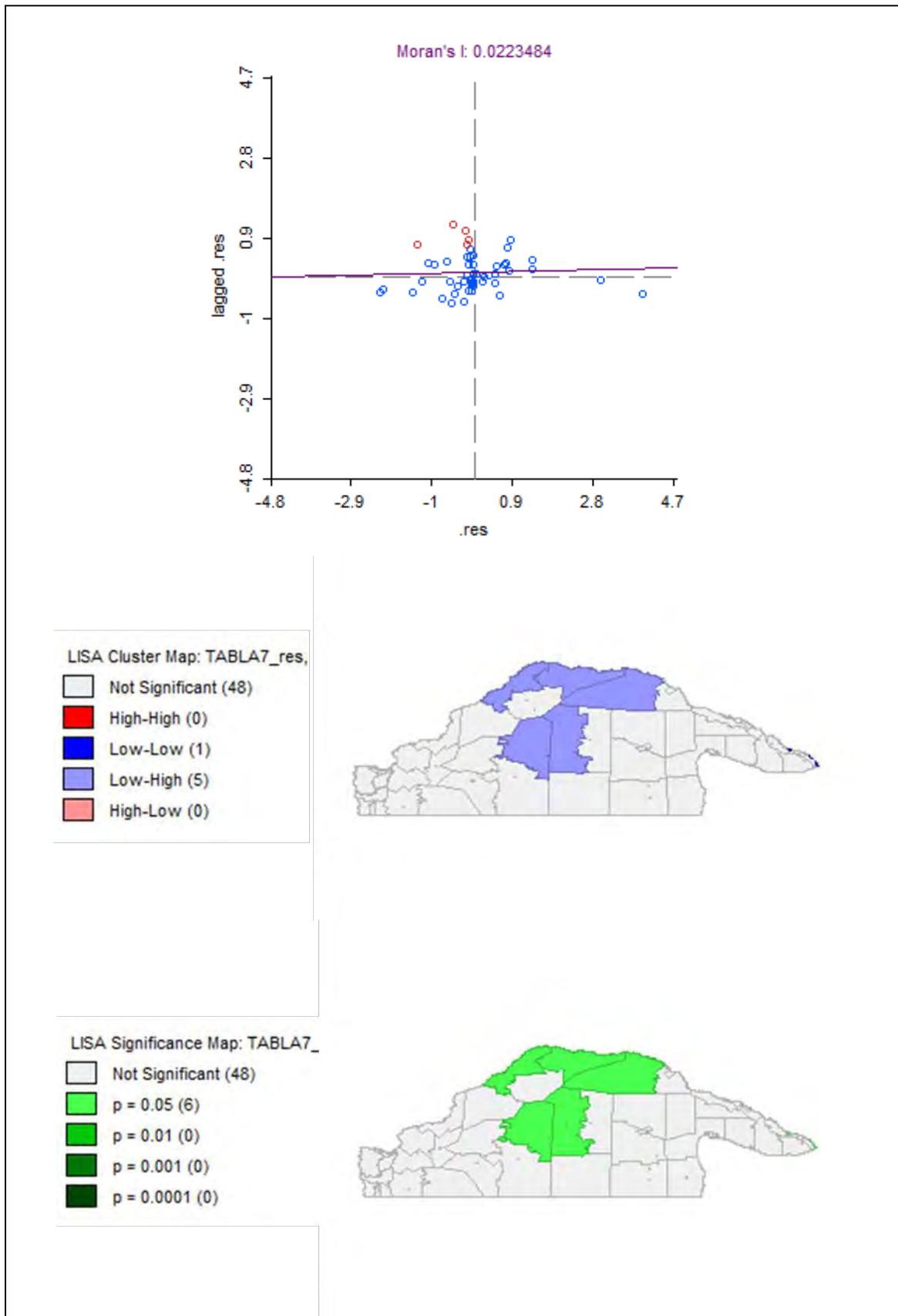


Figura 37. Indicadores locales de asociación espacial para los residuos de MOD\_3.

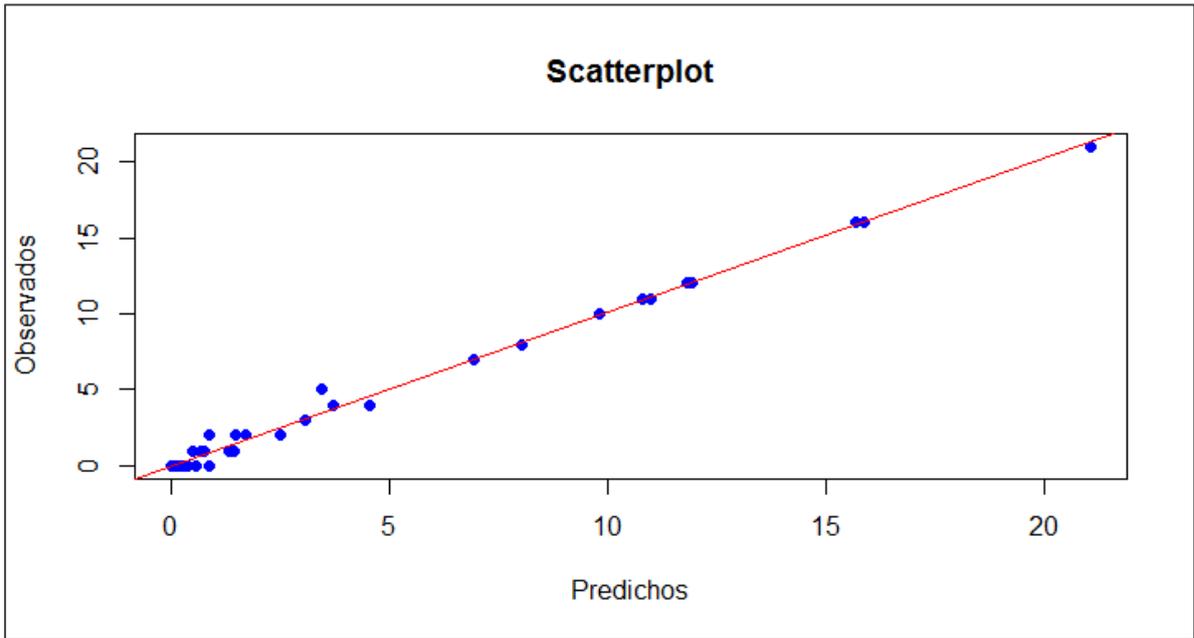


Figura 38. Casos observados vs. Casos predichos por el modelo (MOD\_3).

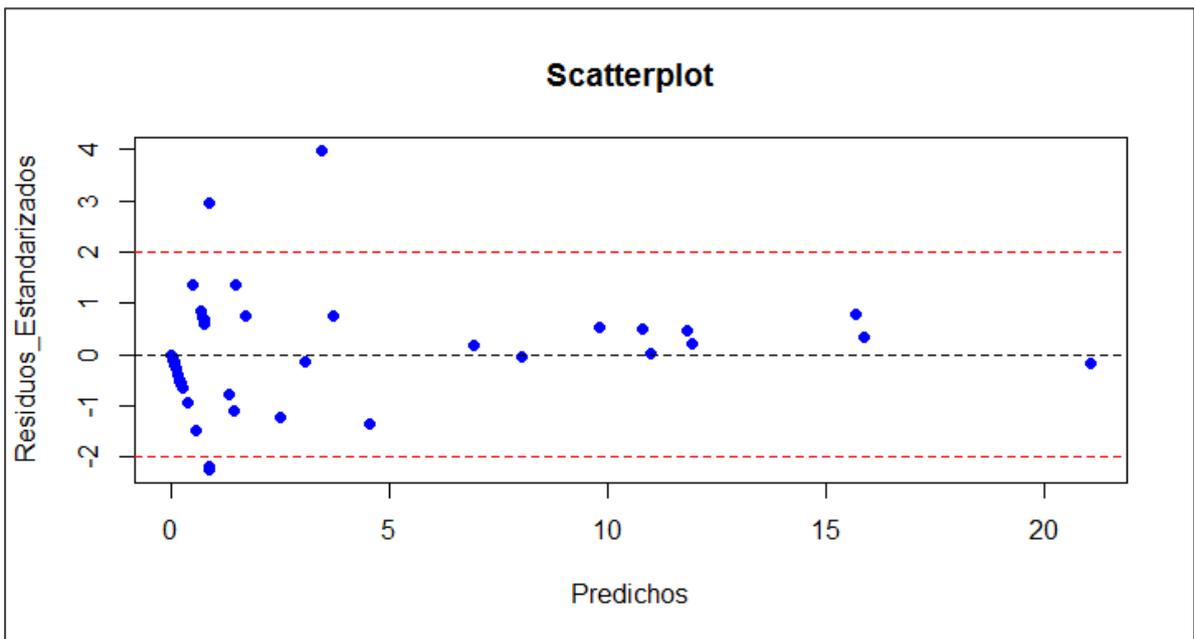


Figura 39. Residuos estandarizados vs. Casos predichos.

## APÉNDICE B: TABLAS

**Tabla 1:**

*Código OpenBUGS para el modelo de covarianza bayesiano jerárquico de efectos fijos con 3 covariables*

---

```
model
{
for (i in 1: region)
{
# verosimilitud de Poisson para los conteos observados
y[i]~dpois(mu[i])
log(mu[i])<-log(E[i])+alpha+(beta1*x1[i]+beta2*x2[i]+beta3*x3[i])
# Riesgo relativo y residuo del riesgo relativo
RR[i]<-exp(alpha+(beta1*x1[i]+beta2*x2[i]+beta3*x3[i]))
res_RR[i]<-exp(alpha)
# Fijados y residuos
fit[i] <- RR[i] * E [i]
res[i] <- y[i] - fit[i]
# Pérdida posterior predictiva (PPL)
ypred[i] ~ dpois(mu[i]) #Distribución predictiva
PPL[i] <- pow(ypred[i]-y[i],2) #Residuos predictivos
PPL2[i] <- abs(ypred[i]-y[i])
}
MAPE <- mean(PPL2[]) # Error absoluto medio predictivo
MSPE <- mean(PPL[]) # Error cuadrático medio predictivo
eps<-1.0E-6
# Distribución a priori impropia para la media del riesgo relativo en la región de estudio
alpha~dflat()
mean<-exp(alpha)
# Distribución a priori de los coeficientes de regresión
beta1~dnorm(0.0,1.0E-5)
beta2~dnorm(0.0,1.0E-5)
beta3~dnorm(0.0,1.0E-5)
}
}
```

---

**Tabla 2:**

**Código Open BUGS para el modelo de covarianza bayesiano jerárquico de efectos mixtos con 3 covariables, con efectos aleatorios no estructurados.**

---

```
model
{
for (i in 1:region)
{
    # verosimilitud de Poisson para los conteos observados
    y[i]~dpois(mu[i])
    log(mu[i])<-log(E[i])+alpha+v[i]+(beta1*x1[i]+beta2*x2[i]+beta3*x3[i])
    # Distribución a priori para heterogeneidad no correlacionada
    v[i]~dnorm(0,tau.v)
    # Riesgo relativo
    RR[i]<-exp(alpha+v[i]+(beta1*x1[i]+beta2*x2[i]+beta3*x3[i]))
    res_RR[i]<-exp(alpha+v[i])
    # Descomposición del riesgo relativo
    RR_ef[i]<-exp(beta1*x1[i]+beta2*x2[i]+beta3*x3[i])
    RR_het[i]<-exp(v[i])
    # Fijados y residuos
    fit[i] <- RR[i] * E [i]
    res[i] <- y[i] - fit[i]
    # Pérdida posterior predictiva (PPL)
    ypred[i] ~ dpois(mu[i]) #Distribución predictiva
    PPL[i] <- pow(ypred[i]-y[i],2) #Residuos predictivos
    PPL2[i] <- abs(ypred[i]-y[i])
}
MAPE <- mean(PPL2[]) # Error absoluto medio predictivo
MSPE <- mean(PPL[]) # Error cuadrático medio predictivo
eps<-1.0E-6
# Distribución a priori impropia para la media del riesgo relativo en la región de estudio
alpha~dflat()
mean<-exp(alpha)
# Distribución a priori de los coeficientes de regresión
beta1~dnorm(0.0,1.0E-5)
beta2~dnorm(0.0,1.0E-5)
beta3~dnorm(0.0,1.0E-5)
# Distribución hiperprior del parámetro inversa de la varianza de los efectos aleatorios
tau.v~dgamma(0.5,0.0005)
# Interpretación del parametro varianza (ej: pow(e1, e2) e1e2):
sd.v <- sd(v[])
var.v <- pow(sd.v,2)
}
```

---

**Tabla 3:**

*Código Open BUGS para el modelo de covarianza bayesiano jerárquico de efectos mixtos con 3 covariables, con efectos aleatorios no estructurados y estructurados.*

---

```
model
{
for (i in 1:region)
{
# verosimilitud de Poisson para los conteos observados
y[i]~dpois(mu[i])
log(mu[i])<-log(E[i])+alpha+v[i]+u[i]+(beta1*x1[i]+beta2*x2[i]+beta3*x3[i])
# Distribución a priori para heterogeneidad no correlacionada
v[i]~dnorm(0,tau.v)
# Riesgo relativo
RR[i]<-exp(alpha+v[i]+u[i]+(beta1*x1[i]+beta2*x2[i]+beta3*x3[i]))
res_RR[i]<-exp(alpha+v[i]+u[i])
# Descomposición del riesgo relativo
RR_ef[i]<-exp(beta1*x1[i]+beta2*x2[i]+beta3*x3[i])
RR_het[i]<-exp(v[i])
RR_clust[i]<-exp(u[i])
# Fijados y residuos
fit[i] <- RR[i] * E [i]
res[i] <- y[i] - fit[i]
# Pérdida posterior predictiva (PPL)
ypred[i] ~ dpois(mu[i]) #Distribución predictiva
PPL[i] <- pow(ypred[i]-y[i],2) #Residuos predictivos
PPL2[i] <- abs(ypred[i]-y[i])
}
MAPE <- mean(PPL2[]) # Error absoluto medio predictivo
MSPE <- mean(PPL[]) # Error cuadrático medio predictivo
eps<-1.0E-6
# CAR distribución a priori para heterogeneidad espacial correlacionada
u[1:m]~car.normal(adj[],weights[],num[],tau.u)
# Weights (ponderación): definido en el código de R
# Distribución a priori impropia para la media del riesgo relativo en la región de estudio
alpha~dflat()
mean<-exp(alpha)
# Distribución a priori de los coeficientes de regresión
beta1~dnorm(0.0,1.0E-5)
beta2~dnorm(0.0,1.0E-5)
beta3~dnorm(0.0,1.0E-5)
# Distribución hiperprior del parámetro inversa de la varianza de los efectos aleatorios
tau.u~dgamma(0.5,0.0005)
tau.v~dgamma(0.5,0.0005)
# Interpretación del parametro varianza (ej: pow(e1, e2) e1e2):
sd.v <- sd(v[])
var.v <- pow(sd.v,2)
sd.u <- sd(u[])
var.u <- pow(sd.u,2)
Alfa <- sd.u/(sd.u+sd.v)
}
}
```

---

**Tabla 4:****Pasos para la selección de covariables con el software R.**

---

```
# Cargar paquetes
> library(sp)
> library(maptools)
# Cargar TABLA7
> TABLA7 <- readShapePoly("E:/Tesis/Capitulos/Capitulo3/TABLAS/TABLA7")
> names(TABLA7)
# Definir variables
> temp <- TABLA7$MediaDiame # Temperatura
> NDVI <- TABLA7$NDVImean # NDVI anual
> PEAO <- TABLA7$DCP_bin # Región natural PEAO
> Precip <- TABLA7$Pre_mean # Precipitación
> Rango_temp <- TABLA7$MediaDia_6 # Rango de temperatura
> SD_temp <- TABLA7$SdDiamean # Desvío estándar de la temperatura anual
> Alt <- TABLA7$Alt_mean # Altura
# Análisis de correlación
> data <- data.frame(temp,Rango_temp,SD_temp,NDVI,PEAO,Precip,SD_NDVI,Alt)
> cor(data)
```

---

**TABLA 5:*****Regresión de Poisson con el software R.***

---

```
# Se cargó el paquete (sp) para clases y métodos de datos espaciales y el paquete (maptools), un
conjunto de herramientas para la manipulación y la lectura de datos geográficos.
> library(sp)
> library(maptools)
# Se cargó desde el directorio de la PC la TABLA7 que contiene las variables a utilizar
> TABLA7 <- readShapePoly("E:/Tesis/Capitulos/Capitulo3/TABLAS/TABLA7")
# Vemos los nombres de las variables
> names(TABLA7)
# Se definieron las variables
> emp <- TABLA7$MediaDiame # Temperatura
> NDVI <- TABLA7$NDVI$mean # NDVI
> PEO <- TABLA7$DCP_bin # Región natural PEO
> Precip <- TABLA7$Pre_mean # Precipitación
# Se calcularon los esperados para cada región
> sum.casos <- sum(TABLA7$Casos)
> sum.pob <- sum(TABLA7$Su_pb_0a14)
> media <- sum.casos/sum.pob
> E <- TABLA7$Su_pb_0a14*media
# Se realizó la regresión de Poisson,
> Model1 <- glm(TABLA7$Casos ~ offset(log(E))+temp+NDVI+PEO+Precip, family=poisson(link =
"log"),data=TABLA7)
> summary(Model1)
```

---

**TABLA 6:**

***Resultados de la regresión de Poisson, procedimiento en el software R: Model1.***

---

```
> Model1 <- glm(TABLA7$Casos ~ offset(log(E))+temp+NDVI+EPAO+Precip, family=poisson(link =  
"log"),data=TABLA7)  
> summary(Model1)
```

Call:

```
glm(formula = TABLA7$Casos ~ offset(log(E)) + temp + NDVI + EPAO +  
Precip, family = poisson(link = "log"), data = TABLA7)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.4492	-0.3179	-0.0896	1.1176	7.3693

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	8.63621	1.72471	5.007	5.52e-07 ***
temp	-0.30825	0.05933	-5.196	2.04e-07 ***
NDVI	-12.89388	2.62309	-4.916	8.85e-07 ***
EPAO	3.45372	0.41211	8.381	< 2e-16 ***
Precip	0.48982	5.31188	0.092	0.927

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1048.00 on 53 degrees of freedom  
Residual deviance: 172.56 on 49 degrees of freedom  
AIC: 263.23

Number of Fisher Scoring iterations: 9

---

**TABLA 7:**

***Resultados de la regresión de Poisson, procedimiento en el software R: Model2.***

---

```
> Model2 <- glm(TABLA7$Casos ~ offset(log(E))+temp+NDVI+EPAO, family=poisson(link =  
"log"),data=TABLA7)  
> summary(Model2)
```

Call:

```
glm(formula = TABLA7$Casos ~ offset(log(E)) + temp + NDVI + EPAO,  
family = poisson(link = "log"), data = TABLA7)
```

Deviance Residuals:

```
Min 1Q Median 3Q Max  
-4.4637 -0.3188 -0.0905 1.1515 7.3729
```

Coefficients:

```
Estimate Std. Error z value Pr(>|z|)  
(Intercept) 8.65777 1.71146 5.059 4.22e-07 ***  
temp -0.30998 0.05633 -5.503 3.73e-08 ***  
NDVI -12.74573 2.07459 -6.144 8.06e-10 ***  
EPAO 3.46001 0.40804 8.480 < 2e-16 ***  
---
```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 1048.00 on 53 degrees of freedom  
Residual deviance: 172.57 on 50 degrees of freedom  
AIC: 261.24
```

Number of Fisher Scoring iterations: 9

---

**TABLA 8:**

*Procedimiento en el software R: Creación del modelo*

---

```
# Declaración del modelo
model {
# Lazo a través de los 54 radios censales
for (i in 1: region) {
# Verosimilitud de Poisson para los conteos observados
# Conteos observados de casos de hidatidosis en cada región (y[i]) siguiendo una distribución de Poisson
(dpois) con parámetro lambda para cada región (que es la media) igual a mu[i]
  y[i] ~ dpois(mu[i])
# Logaritmo de mu[i] es igual al logaritmo de los conteos de hidatidosis esperados más theta[i]
  log(mu[i])<-log(E[i]) + theta[i]
# theta[i] es igual al intercepto (alpha) más el coeficiente de regresión ( $\beta_m$ ) por el valor de la covariable
en cada región ( $x_m[i]$ )
  theta[i] <- alpha+(beta1*x1[i]+beta2*x2[i]+beta3*x3[i])
# El riesgo relativo RR[i] para cada región corresponde al exponencial de theta[i]
  RR[i]<-exp(theta[i])
# El conteo predicho fit[i] está dado por el producto del riesgo relativo RR[i] por los esperados E[i], y los
residuos son los observados y[i] menos los predichos fit[i] para cada región.
  fit[i] <- RR[i] * E [i]
  res[i] <- y[i] - fit[i]
}
# La distribución a priori impropia para la media del riesgo relativo en la región de estudio. La
distribución a priori para alpha es normal (dnorm) con media cero y precisión (inversa a la varianza) de
1.0E-5.
alpha ~ dnorm(0.0,1.0E-5)
mean <- exp(alpha)
# La distribución a priori para los coeficientes de regresión es normal (dnorm) con media cero y
precisión (inversa a la varianza) de 1.0E-5.
beta1~dnorm(0.0,1.0E-5)
beta2~dnorm(0.0,1.0E-5)
beta3~dnorm(0.0,1.0E-5)
}
```

---

**TABLA 9:**

*Procedimiento en el software R: datos y parámetros*

---

```
# Se cargó la librería para la inferencia Bayesiana
> library(R2OpenBUGS)
# Se definieron los datos como el número de regiones, los casos, los esperados y las covariables.
> region <- nrow(TABLA7)
> data1 <- list(y = TABLA7$Casos, E = E, x1 = TABLA7$MediaDiame , x2 = TABLA7$NDVImean, x3 =
TABLA7$DCP_bin, region = region)
# Al asignar los valores iniciales se tuvo en cuenta que en modelos complejos WinBUGS necesita
comenzar las cadenas de Markov no demasiado lejos de su distribución estacionaria o se detiene o ni
siquiera empezara a actualizar.
# El requisito de iniciar las cadenas cercanas a la solución va en contra del requisito de iniciar éstas en
lugares dispersos con el fin de evaluar la convergencia, por lo que se tomó una elección intermedia
razonable (Kéry, 2010).
> inits1 <- list(alpha=0, beta1=0, beta2=0, beta3=0)
> inits2 <- list(alpha=8.7, beta1=-0.3, beta2=-13, beta3=3.4)
> inits3 <- list(alpha=17, beta1=-0.6, beta2=-26, beta3=7)
> inits <- list(inits1,inits2,inits3)
# Se indicó los parámetros a estimarse, los que se definen dentro del modelo.
> params <- c("alpha", "beta1", "beta2", "beta3", "RR", "res_RR", "RR_ef", "fit", "res",
"PP", "res_PP", "mean", "MAPE", "MSPE")
# Para la configuración del método de Cadenas de Markov Monte Carlo, se corrieron tres cadenas (nc),
a fin de evaluar la performance del modelo con diferentes valores iniciales. Se estableció el número de
muestras de la distribución a posteriori por cadena (ni) y el número de muestras a descartar (nb), que es
la porción inicial no estacionaria de la cadena, conocida como el "burn-in" (Meyer, 2014). Por último se
especifica la tasa (nt) con que se almacenaran las muestras desde cada k-esima iteración, denominada
"Thinning rate" según el manual de usuario de WinBUGS.
> nc <- 3
> ni <- 50000
> nb <- 25000
> nt <- 5
# El muestreo de Gibbs se comenzó alimentando la función bugs() con los datos, los valores iniciales, los
parámetros, el directorio donde se guardó el modelo (que declaramos previamente), el número de
cadenas, el número de iteraciones, el número de iteraciones para descartar al comienzo y la tasa con
que se almacenarán las muestras. La opción debug= TRUE significa que OpenBUGS permanecerá
abierto para futuras investigaciones.
> out <- bugs(data=data1, inits=inits, parameters.to.save=params, model.file =
"E:/Tesis/Capitulos/T1N1E4/Ecological_analisis_Poisson_efectos_fijos3.txt", n.thin=nt, n.chains=nc,
n.burnin=nb, n.iter=ni, debug = TRUE)
# Luego de terminado el proceso en OpenBUGS se imprimen y guardan los resultados
> print(out, dig = 3)
> save(out, file="E:/Tesis/Capitulos/T1N1E4/outEF.RData")
```

---

**TABLA 10:**

*Comparación de los coeficientes de regresión de los modelos de EF computados usando estadística frecuentista (en R) y Bayesiana (en OpenBUGS).*

Método	Alpha	Temp	NDVI	PEAO
Frecuentista	8.6362	-0.3082	-12.8939	3.4537
Bayesiano completo	8.6170	-0.3107	-12.7500	3.5080

**TABLA 11:**

*Resumen de estadísticos del modelo de EF Bayesiano.*

Modelo	Variable	Media Posterior	SD	MC error	RR	RR (IC del 95%)
MOD_1	Intercepto	8.6170	1.7110	0.0153		
	temperatura	-0.3107	0.0565	0.0005	0.7329	0.6568 - 0.8181
	NDVI	-12.7500	2.0720	0.0180	0.0000	0.0000 - 0.0002
	PEAO	3.5080	0.4074	0.0037	33.3814	15.3023 - 76.0203

**TABLA 12:**

*Procedimiento en el software R para el cálculo del estadístico Moran's I para los residuos.*

---

```
# Para el modelo de efectos fijos, se crea una matriz de adyacencia espacial tipo reina, es decir,
# polígonos teniendo un borde en común o esquina se ponderaron con 1 o 0 de otro modo.
> library(maps)
> library(spdep)
> TABLA7.nb <- poly2nb(TABLA7) #adjacency matrix
> TABLA7.wts <- nb2listw(TABLA7.nb, zero.policy = T, style = "B")
> print(TABLA7.wts, zero.policy = T)
> names(TABLA7.wts)
# Para evaluar la autocorrelación en los residuos se calcula el índice de Moran's I y test de significancia.
> Library(spdep)
> x <- data$res
> residuos <- as.vector(x)
> moran.test(residuos,TABLA7.wts,randomisation=TRUE,zero.policy=TRUE)
# Se grafican los resultados
> moran.plot(residuos,TABLA7.wts,zero.policy=TRUE,labels=NULL,xlim=c(-4,4) )
```

---

**TABLA 13:**

*Procedimiento en R para el estadístico de Moran's I para cuantificar la autocorrelación espacial en áreas.*

---

```
> moran.test(residuos,TABLA7.wts,randomisation=TRUE,zero.policy=TRUE)
```

Moran's I test under randomisation

data: residuos

weights: TABLA7.wts

Moran I statistic standard deviate = -1.7892, p-value = 0.9632

alternative hypothesis: greater

sample estimates:

Moran I statistic	Expectation	Variance
-0.158059458	-0.018867925	0.006051822

Moran's I scatterplot

```
> moran.plot(residuos,TABLA7.wts,zero.policy=TRUE,labels=NULL,xlim=c(-4,4) )
```

---

**Tabla 14:**

*Procedimiento en el software R: Creación del modelo (MOD\_3: modelo de EM con efectos aleatorios estructurados y no estructurados).*

---

```
# Declaración del modelo: Modelo de efectos Mixtos
model {
# Lazo a través de los 54 radios censales
for (i in 1: region) {
# Verosimilitud de Poisson para los conteos observados
y[i]~dpois(mu[i])
log(mu[i])<-log(E[i]) + alpha+v[i]+u[i]+(beta1*x1[i]+beta2*x2[i]+beta3*x3[i])
# Riesgo Relativo
RR[i]<-exp(alpha+v[i]+u[i]+(beta1*x1[i]+beta2*x2[i]+beta3*x3[i]))
res_RR[i]<-exp(alpha+v[i]+u[i])
# Descomposición del Riesgo Relativo
RR_ef[i]<-exp(beta1*x1[i]+beta2*x2[i]+beta3*x3[i])
RR_het[i]<-exp(v[i])
RR_clust[i]<-exp(u[i])
# Predichos y residuos
fit[i] <- RR[i] * E [i]
res[i] <- y[i] - fit[i]
# Probabilidad posterior de RR[i]>1
PP[i]<-step(RR[i]-1+eps)
res_PP[i]<-step(res_RR[i]-1+eps)
# Distribución a priori para la heterogeneidad no correlacionada
v[i]~dnorm(0,tau.v)
# Perdida posterior predictiva (PPL)
ypred[i] ~ dpois(mu[i]) #Distribución predictiva
PPL[i] <- pow(ypred[i]-y[i],2) #Residuos predictivos
PPL2[i] <- abs(ypred[i]-y[i])
}
MAPE <- mean(PPL2[]) # error absoluto
MSPE <- mean(PPL[]) # Error predictivo cuadrado medio
eps<-1.0E-6
# CAR distribución a priori para la heterogeneidad correlacionada.
u[1:m]~car.normal(adj[],weights[],num[],tau.u)
# Weights: definido en el código de R
# Distribución a priori inpropia para la media del Riesgo Relativo de la región de estudio
alpha~dflat()
mean<-exp(alpha)
# Distribución a priori de los coeficientes de regresión
beta1~dnorm(0.0,1.0E-5)
beta2~dnorm(0.0,1.0E-5)
beta3~dnorm(0.0,1.0E-5)
# Distribución hiperprior del parámetro varianza inversa de los efectos aleatorios
tau.u~dgamma(0.5,0.0005)
tau.v~dgamma(0.5,0.0005)
# Interpretación del parámetro varianza (ej: pow(e1, e2) e1e2):
```

```
sd.v <- sd(v[])  
var.v <- pow(sd.v,2)  
sd.u <- sd(u[])  
var.u <- pow(sd.u,2)  
Alfa <- sd.u/(sd.u+sd.v) # Relación entre los componentes de la heterogeneidad  
}
```

---

**Tabla 15:**

***Procedimiento en el software R: datos y parámetros.***

---

```
# Se cargó la librería para la inferencia Bayesiana
> library(R2OpenBUGS)
# Se definieron los datos como el número de regiones, los casos, los esperados y las covariables.
> region <- nrow(TABLA7)
> data1 <- list(y = TABLA7$Casos, E = E, x1 = TABLA7$MediaDiame, x2 = TABLA7$NDVImean, x3 =
TABLA7$DCP_bin, region = region)
# Se asignaron los valores iniciales
> inits1 <- list(alpha=0, beta1=0, beta2=0, beta3=0)
> inits2 <- list(alpha=8.7, beta1=-0.3, beta2=-13, beta3=3.4)
> inits3 <- list(alpha=17, beta1=-0.6, beta2=-26, beta3=7)
> inits <- list(inits1,inits2,inits3)
# Se indicó los parámetros a estimarse, los que se definen dentro del modelo.
> Params<-c("alpha","beta1","beta2","beta3","sd.v","sd.u","MAPE","MSPE",
"RR","res_RR","PP","res_PP","RR_ef","RR_het","RR_clust","Alfa")
# Configuración del método, Cadenas de Markov Monte Carlo.
nc <- 3 # Número de cadenas
ni <- 150000 # Número de muestras de la distribución a posteriori por cadena
nb <- 75000 # Número de muestras a descartar
nt <- 10 # Tasa con que se almacenaran las muestras desde cada k-esima iteración
# Se comienza el muestreo de Gibbs
> out_EM<- bugs(data= c(data1,TABLA7.WBweights), inits=inits, parameters.to.save=params,
model.file="E:/Tesis/Capitulos/Capitulo4/Modelos_WinBUGS3/TABLA7/T1N1E4/Ecological_analysis_Pois
son_efectos_mixtos_3.txt", n.thin=nt, n.chains=nc, n.burnin=nb, n.iter=ni, debug = TRUE)#,
codaPkg=TRUE)
# Luego de terminado el proceso en OpeeBUGS se imprimieron y guardaron los resultados
> print(out, dig = 3)
> save(out_EM, file="E:/Tesis/Capitulos/T1N1E4/out.EM.RData")
# Se evaluó la salida con el paquete CODA
> out.CODA <- read.bugs(out_EM)
> library(coda)
> str(out.CODA)
> varnames(out.CODA)
> plot(out.CODA[,55:56])
> autocorr.plot(out.CODA[,55:56])
> gelman.plot(out.CODA[,55:56])
> gelman.diag(out.CODA[,55:56])
> summary(out.CODA[,55:56])
```

---

**Tabla 16:**

*Procedimiento en el software R: estandarización y centrado de variables.*

---

```
> TABLA7$MediaDiame.R <- rescale(TABLA7$MediaDiame,"full")
> TABLA7$NDVImean.R <- rescale(TABLA7$NDVImean,"full")
> TABLA7$DCP_bin.R <- rescale(TABLA7$DCP_bin, binary.inputs="center")
```

---

**Tabla 17:**

*Resultados de los modelos de EF (MOD\_1), modelo de EM con efectos aleatorios no estructurados (MOD\_2) y modelo de EM con efectos aleatorios estructurados y no estructurados (MOD\_3).*

Modelo	Variable	Media Posterior	SD	MC error	RR	RR (IC del 95%)
MOD_1	Intercepto	8.6170	1.7110	0.0153		
	Temperatura	-0.3107	0.0565	0.0005	0.7329	0.6568 - 0.8181
	NDVI	-12.7500	2.0720	0.0180	0.0000	0.0000 - 0.0002
	PEAO	3.5080	0.4074	0.0037	33.3814	15.3023 - 76.0203
MOD_2	Intercepto	9.8760	3.8470	0.0356		
	Temperatura	-0.3038	0.1221	0.0011	0.7380	0.5729 - 0.9282
	NDVI	-14.7200	5.1630	0.0510	0.0000	0.0000 - 0.0075
	PEAO	2.3890	0.9765	0.0100	10.9026	1.7186 - 80.3185
	heterogeneidad * No estructurada	1.8500	0.3037	0.0024		
MOD_3	Intercepto	9.5790	4.0250	0.0298		
	Temperatura	-0.2905	0.1325	0.0012	0.7479	0.5756 - 0.9722
	NDVI	-14.6000	5.2570	0.0303	0.0000	0.0000 - 0.0112
	PEAO	2.3070	1.0260	0.0091	10.0442	1.3623 - 79.9179
	heterogeneidad * No estructurada	1.7990	0.3669	0.0047		
	* Estructurada	0.2230	0.5064	0.0122		
	Alfa	0.0900				

**Tabla 18:**

*Bondad de ajuste para los diferentes modelos.*

<b>Modelo</b>	<b>MSPE</b>	<b>Desviación</b>	<b>pd</b>	<b>DIC</b>
MOD_1	14.26	257.3	3.97	261.2
MOD_2	5.76	119.8	26.1	145.9
MOD_3	5.77	119.8	18.96	138.7

**Tabla 19:**

*Coefficientes de regresión para las covariables estandarizadas y centradas, y varianza de los componentes aleatorios no estructurados y estructurados (sd.u y sd.s respectivamente) de los diferentes modelos.*

Modelo	Variable (2 sd's)	Media Posterior	SD	MC error	RR	RR (IC del 95%)
MOD_1	Intercepto	-0.7267	0.1999	0.0009		
	temperatura	-3.1590	0.5764	0.0024	0.0425	0.0137 - 0.1304
	NDVI	-3.3560	0.5485	0.0022	0.0349	0.0119 - 0.1016
	PEAO	3.5020	0.4117	0.0018	33.1817	15.2107 - 76.7075
MOD_2	Intercepto	-0.2732	0.4706	0.0045		
	temperatura	-3.0740	1.2350	0.0105	0.0462	0.0036 - 0.4669
	NDVI	-3.8670	1.3600	0.0126	0.0209	0.0013 - 0.2822
	PEAO	2.3910	0.9814	0.0107	10.9244	1.7045 - 82.5166
	heterogeneidad * No estructurada	1.8530	0.3070	0.0025		
MOD_3	Intercepto	-0.2718	0.4686	0.0038		
	temperatura	-2.9420	1.3750	0.0182	0.0528	0.0036 - 0.8178
	NDVI	-3.8570	1.3970	0.0119	0.0211	0.0013 - 0.3107
	PEAO	2.2910	1.0290	0.0128	9.8848	1.3461 - 80.1580
	heterogeneidad * No estructurada	1.7920	0.3708	0.0060		
	* Estructurada	0.2403	0.5493	0.0173		
	Alfa	0.0900				

**Tabla 20:**

*Modelo de EF y EM, usando como covariable la temperatura media anual de la superficie terrestre en forma binaria (valor = 1, para los radios censales que están en el rango de temperatura de 17.19 a 20.90 °C y valor = 0, para el resto).*

Modelo	Variable	Media Posterior	SD	MC error	RR	RR (IC del 95%)	DIC*
Modelo de efectos fijos							
Temperatura media anual (Rango óptimo)	beta0	-1.10	0.14	0.0005			527.09
	beta1	4.54	0.17	0.0007	<b>93.78</b>	67.29 - 132.03	
Modelo de EM							
Temperatura media anual (Rango óptimo)	beta0	-0.64	0.50	0.0048			166.37
	beta1	2.42	1.02	0.0229	<b>11.23</b>	1.68 - 100.79	
	sd.s	2.26	0.90	0.0285			
	sd.u	1.33	0.65	0.0173			
	Alfa	0.63					

**Nota.** DIC\* computado, con el número efectivo de parámetros (pD) de la forma  $pD = \text{Var}(D) / 2$ .

**Tabla 21:**

*Análisis de sensibilidad usando diferentes valores para la distribución Gamma del parámetro varianza inversa de los efectos aleatorios.*

Hiperprior parámetros				
	(0.01, 0.01)	(0.001, 0.001)	(0.5, 0.0005)	(0.1, 0.1)
Parámetros	media	media	media	media
Alpha	-0.40	-0.38	-0.27	-0.41
beta1	-2.31	-2.43	-2.96	-2.17
beta2	-3.78	-3.78	-3.84	-3.75
beta3	1.85	1.94	2.31	1.76
sd.s	1.27	1.09	0.21	1.44
sd.u	1.60	1.65	1.81	1.56
DIC*	160.35	160.63	162.80	160.69

**Nota.** DIC\* computado, con  $pD$  de la forma  $pD = \text{Var}(D) / 2$ .

**Tabla 22:**

*Resultados del test de Moran's I para el modelo de EM con efectos aleatorios incorporados (MOD\_3).*

Moran's I test under randomisation		
data: residuos		
weights: TABLA7.wts		
Moran I statistic standard deviate = -0.081935, p-value = 0.5327		
alternative hypothesis: greater		
sample estimates:		
Moran I statistic	Expectation	Variance
-0.025703845	-0.018867925	0.006960691
Moran's I scatterplot		
> moran.plot(residuos,TABLA7.wts,zero.policy=TRUE,labels=NULL,xlim=c(-4,4) )		

## APÉNDICE C: USO COMBINADO DE R Y OPENBUGS: BREVE DESCRIPCIÓN

---

El proyecto BUGS (inferencia Bayesiana Usando muestreo de Gibbs) se ocupa de software flexibles para el análisis bayesiano de modelos estadísticos complejos utilizando MCMC. Dichos softwares son WinBUGS y su primo de código abierto OpenBUGS, a los que se puede acceder desde la sección software de la página web de la unidad de bioestadística del consejo médico de investigación “MRC Biostatistics Unit” (MRC Biostatistics Unit, 2016), donde además puede encontrarse una variedad de recursos como bibliografía por ejemplo. WinBUGS es un programa que como su nombre lo indica es solo para Window. Posee una forma de escribir scripts similar al lenguaje de programación de R y tiene una interfaz gráfica basada en grafos dirigidos acíclicos llamado "Doodles" donde se puede llamar al modelo. OpenBUGS funciona muy parecido a WinBUGS.

Otro programa independiente es justo otro muestreo de Gibbs (JAGS), que también ejecuta análisis bayesiano. También hay un creciente número de paquetes y funciones de R para hacer análisis bayesiano. Para ver estos uno puede dirigirse a la vista de tareas de CRAN para la Inferencia Bayesiana (Park, 2015)

Se puede llamar a WinBUGS, OpenBUGS y JAGS desde dentro de R y además se pueden crear R objetos que se integran perfectamente con varias herramientas espaciales. Aquí se trabaja de esa manera.

Primero se debe instalar OpenBUGS en la computadora (PC). Entonces se escribe un modelo en un archivo de texto y se guarda en alguna parte de la PC. A continuación, se hace una conexión con el programa OpenBUGS desde R utilizando una función del paquete R2WinBUGS () que le dice a OpenBUGS donde se encuentra el modelo en la PC, cuáles son los datos, y qué parámetros definimos para las variables. Si todo va bien, R se conecta a OpenBUGS, suministrando la información que ya se ha proporcionado a la función de R, se ejecuta la simulación, y esta devuelve los resultados como un objeto que puede ser analizado en R.

---

## APÉNDICE D: PROCEDIMIENTO EN EL SOFTWARE R: EVALUACIÓN DE LA SALIDA DEL MUESTREO DE GIBBS CON CODA.

---

Si bien se pueden analizar los resultados de la ejecución de BUGS en su totalidad con las herramientas disponibles en R, se utilizó un par de paquetes diseñados específicamente para esta tarea, CODA (diagnóstico de convergencia y análisis de la salida) y BOA (análisis de salida Bayesiana). Estos paquetes básicamente hacen las mismas cosas, aquí se trabajó con el paquete CODA. El paquete CODA requiere un archivo especialmente estructurado llamado “objeto mcmc”. Conseguir este archivo es un proceso de dos pasos en R2WinBUGS. En primer lugar, se ejecutó la simulación como se haría normalmente, excepto que se incluye la opción `codaPkg = TRUE`. Esto crea un archivo de texto para cada cadena. Luego se ejecutó el comando `read.bugs ()` sobre el objeto `bugs` resultante para crear el objeto `mcmc` que el paquete `coda` requiere.

```
# Entonces se realizó el muestreo de Gibbs con la opción codaPkg = TRUE

> out <- bugs(data=data1, inits=inits, parameters.to.save=params, model.file =
"E:/Tesis/Capitulos/Capitulo4/Modelos_WinBUGS3/TABLA7/T1N1E4/Ecological_analisi
s_Poisson_efectos_fijos3.txt", n.thin=nt, n.chains=nc, n.burnin=nb, n.iter=ni, debug =
TRUE, codaPkg = TRUE)

# Se lee el objeto bugs de la salida del muestreo de Gibbs y se vuelve a y guardar los
resultados

> out.CODA <- read.bugs(out)

> save(out.CODA, file="E:/Tesis/Capitulos/T1N1E4/outEF_CODA.RData")

# Se carga la librería CODA y procedemos con la inspección de los resultados

> library(coda)

# Se observa la estructura del objeto mcmc (FIGURA D.1).

> str(out.CODA)
```

El objeto `mcmc` es una lista que consta de 3 cadenas, cada cadena tiene 25000 x 386 dimensiones correspondientes a 25000 repeticiones de 386 variables (recordemos que hay 54 radios censales y cada uno tiene un riesgo relativo, residuos, etc.). Los comandos de CODA dan resultados para cada una de esas 386 variables. Para esta demostración, se extrae sólo un pequeño número de esas variables. Sin embargo, se deben chequear todos los resultados.

Aunque esta es una lista, hay un método directo para seleccionar un sub-set de objetos MCMC (`[i, j]`), donde `i` es la fila a extraer y `j` es la columna. Se utiliza la función `varnames()` para detectar que columnas corresponden a las cadenas de las estimaciones que se quieren evaluar (FIGURA D.2).

```

List of 3
 $ : mcmc [1:25000, 1:386] 2.11 1.94 2 1.98 1.94 ...
  ..- attr(*, "dimnames")=List of 2
  ...$: chr [1:25000] "25001" "25002" "25003" "25004" ...
  ...$: chr [1:386] "MAPE" "MSPE" "PP[1]" "PP[2]" ...
  ..- attr(*, "mcpair")= num [1:3] 25001 50000 1
 $ : mcmc [1:25000, 1:386] 1.98 2.26 1.59 1.68 2.39 ...
  ..- attr(*, "dimnames")=List of 2
  ...$: chr [1:25000] "25001" "25002" "25003" "25004" ...
  ...$: chr [1:386] "MAPE" "MSPE" "PP[1]" "PP[2]" ...
  ..- attr(*, "mcpair")= num [1:3] 25001 50000 1
 $ : mcmc [1:25000, 1:386] 1.65 2.3 1.94 1.63 1.46 ...
  ..- attr(*, "dimnames")=List of 2
  ...$: chr [1:25000] "25001" "25002" "25003" "25004" ...
  ...$: chr [1:386] "MAPE" "MSPE" "PP[1]" "PP[2]" ...
  ..- attr(*, "mcpair")= num [1:3] 25001 50000 1
 - attr(*, "class")= chr "mcmc.list"

```

**Figura D.1. Estructura del objeto mcmc.**

```
> varnames(out.CODA)
```

```

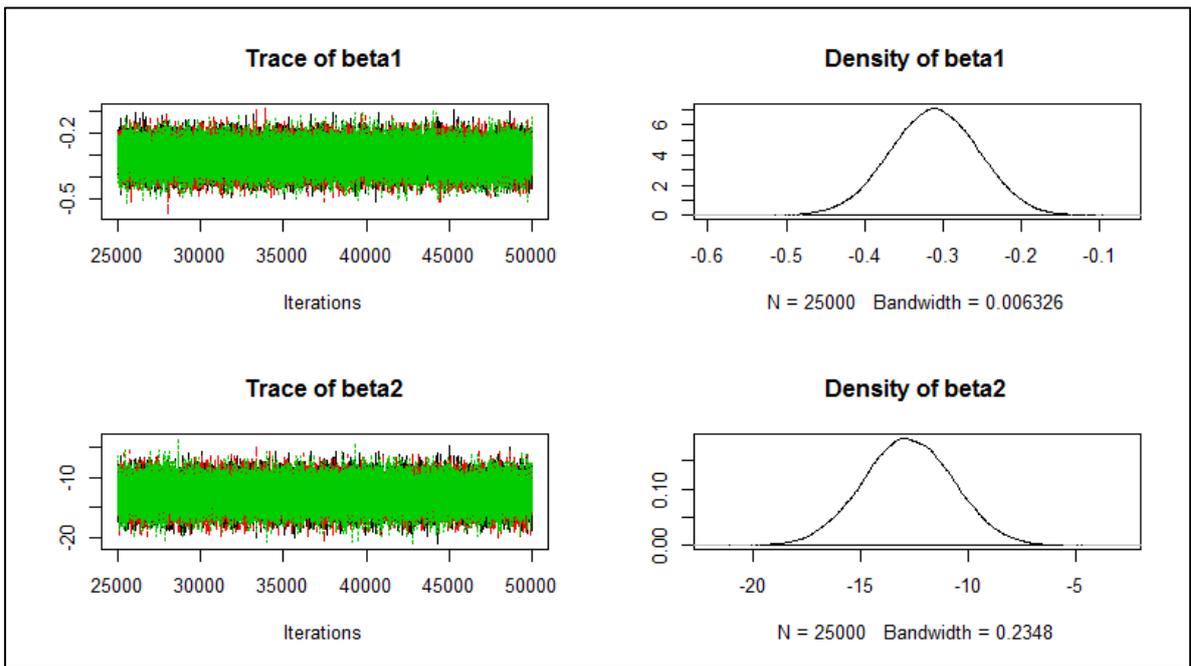
 [1] "MAPE"      "MSPE"      "PP[1]"     "PP[2]"     "PP[3]"
 [6] "PP[4]"     "PP[5]"     "PP[6]"     "PP[7]"     "PP[8]"
[11] "PP[9]"     "PP[10]"    "PP[11]"    "PP[12]"    "PP[13]"
[16] "PP[14]"    "PP[15]"    "PP[16]"    "PP[17]"    "PP[18]"
[21] "PP[19]"    "PP[20]"    "PP[21]"    "PP[22]"    "PP[23]"
[26] "PP[24]"    "PP[25]"    "PP[26]"    "PP[27]"    "PP[28]"
.....

```

**FIGURA D.2. Nombres de la salida de CODA**

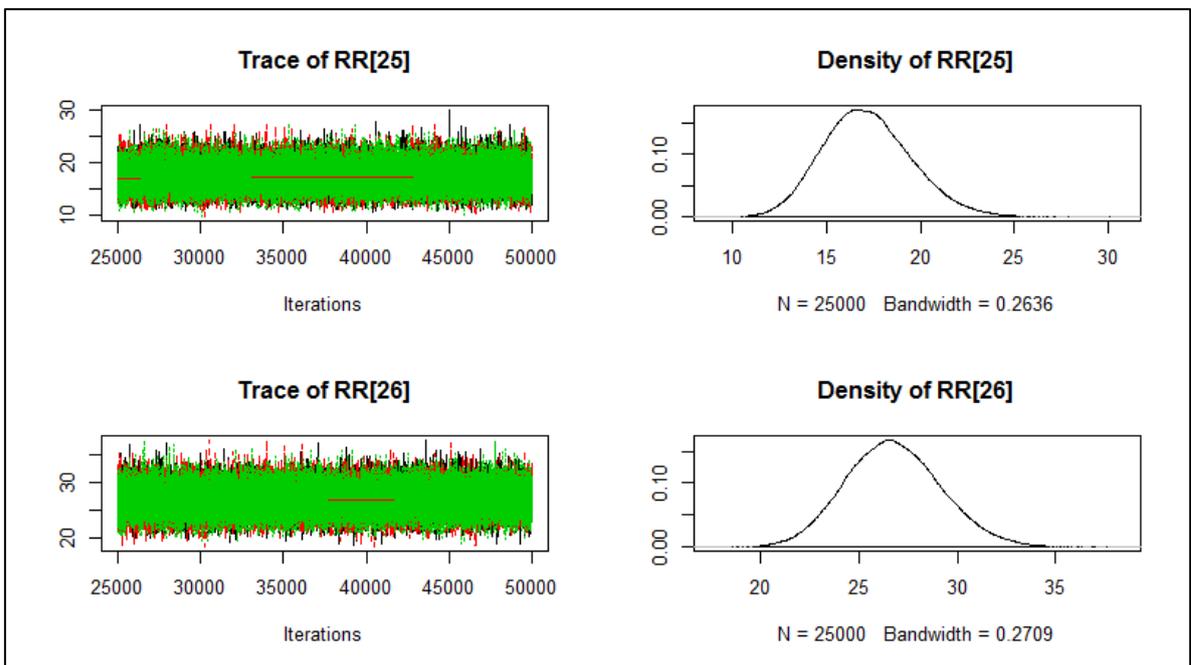
Se trazaron las cadenas como series de tiempo y se inspeccionaron para ver si se superponen y convergen a una distribución estable. Se evaluó su apariencia para ver si parecen razonables, se solapan de manera aceptable "toman forma de oruga", no parecen "vagar" y se "establecen" o convergen a un patrón estable. Mediante la función plot() se observan las densidades de las cadenas, evaluando si parecen estar normalmente distribuida alrededor de una media que podemos utilizar como un valor de resumen de los coeficientes o para el mapeo. Como ejemplo graficamos los coeficientes para las dos primeras covariables del modelo de EF (temperatura y NDVI) (FIGURA D.3) y para el RR de los radios censales 25 y 26 (FIGURA D.4).

```
> plot(out.CODA[,166:167])
```



**FIGURA D.3.** Traza y densidad de distribución de los coeficientes de la temperatura media anual ( $\beta_1$ ) y NDVI medio anual ( $\beta_2$ ).

```
> plot(out.CODA[,81:82])
```

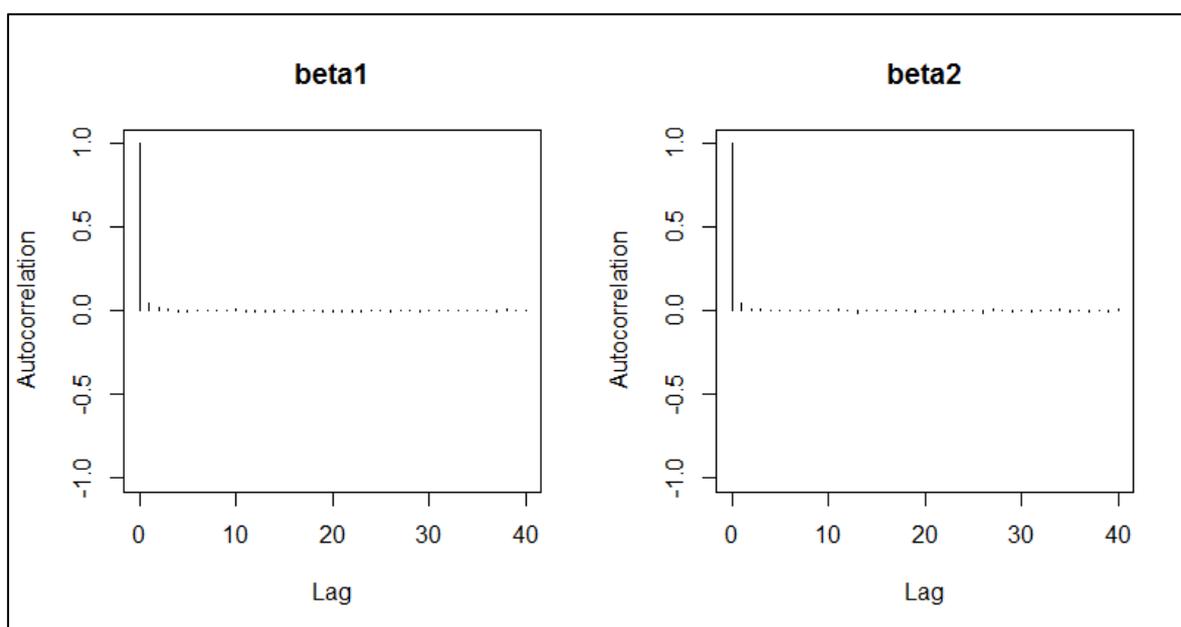


**FIGURA D.4.** Traza y densidad de distribución de los RR de los radios censales 25 y 26.

Se tuvo en cuenta que las densidades graficadas en el lado derecho de la figura representan la distribución de probabilidad real de las estimaciones de riesgo relativo, y que conceptualmente esto es muy diferente de la estadística frecuentista.

Con el paquete CODA se trazaron fácilmente las autocorrelaciones de las cadenas. Por definición, cada paso en una cadena MCMC se basa en el paso inmediato anterior, por lo que las simulaciones están correlacionadas. Por ello se evaluó que estas autocorrelaciones disminuyan a medida que la simulación atraviesa el espacio de probabilidad y se instala la convergencia. Se puede ver en el siguiente gráfico (FIGURA D.5 Y D.6) cómo la autocorrelación disminuye rápida y precipitadamente, lo que indica una convergencia adecuada.

```
> autocorr.plot(out.CODA[,166:167])
```

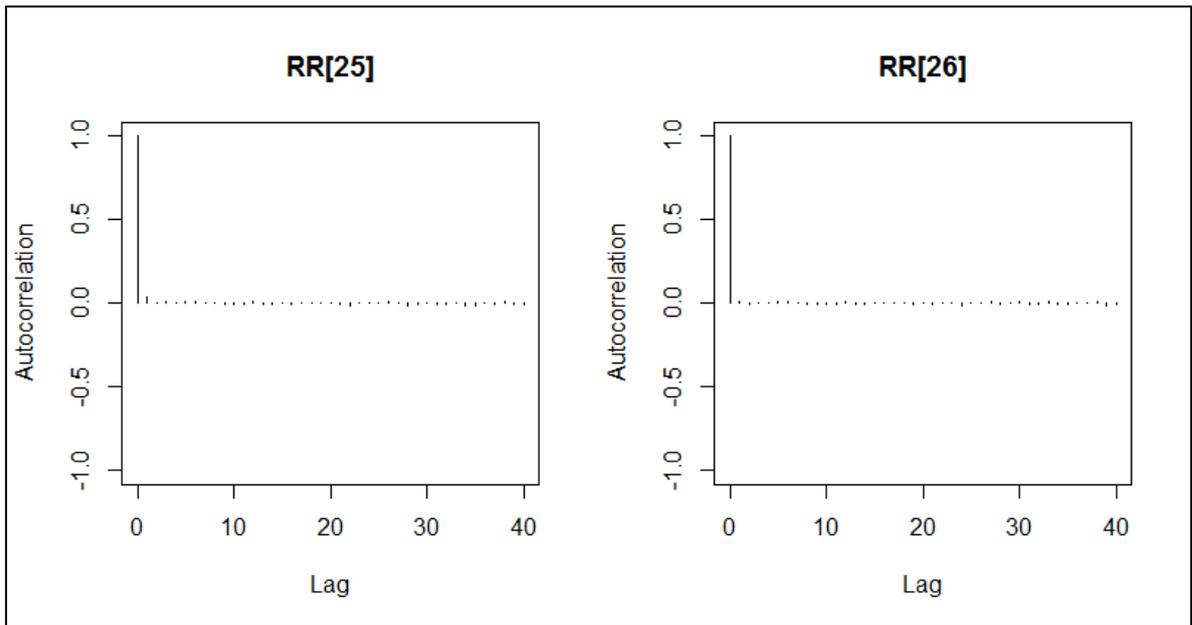


**FIGURA D.5. Grafica de la autocorrelación de las cadenas para la temperatura media anual (beta1) y NDVI medio anual (beta2).**

Como la convergencia de una simulación MCMC es un tema crítico en el análisis bayesiano, se inspeccionó ésta gráficamente y con pruebas estadísticas como Gelman-Rubin que están disponibles para ayudar a evaluar mejor el algoritmo.

El estadístico Gelman-Rubin (Gelman y Rubin, 1992; Brooks y Gelman, 1998) se basa en un enfoque similar a ANOVA para las medias de dos o más cadenas. Un valor cercano a 1 indica la convergencia. Coda viene con versión gráfica y una estadística del diagnóstico de Gelman-Rubin. Se realizan ambas cosas (FIGURA D.7 Y D.8).

```
> autocorr.plot(out.CODA[,81:82])
```



**FIGURA D.6. Grafica de la autocorrelación del RR de los radios censales 25 y 26.**

```
> gelman.diag(out.CODA[,81:82])

Potential scale reduction factors:

      Point est. Upper C.I.
RR[25]      1      1
RR[26]      1      1

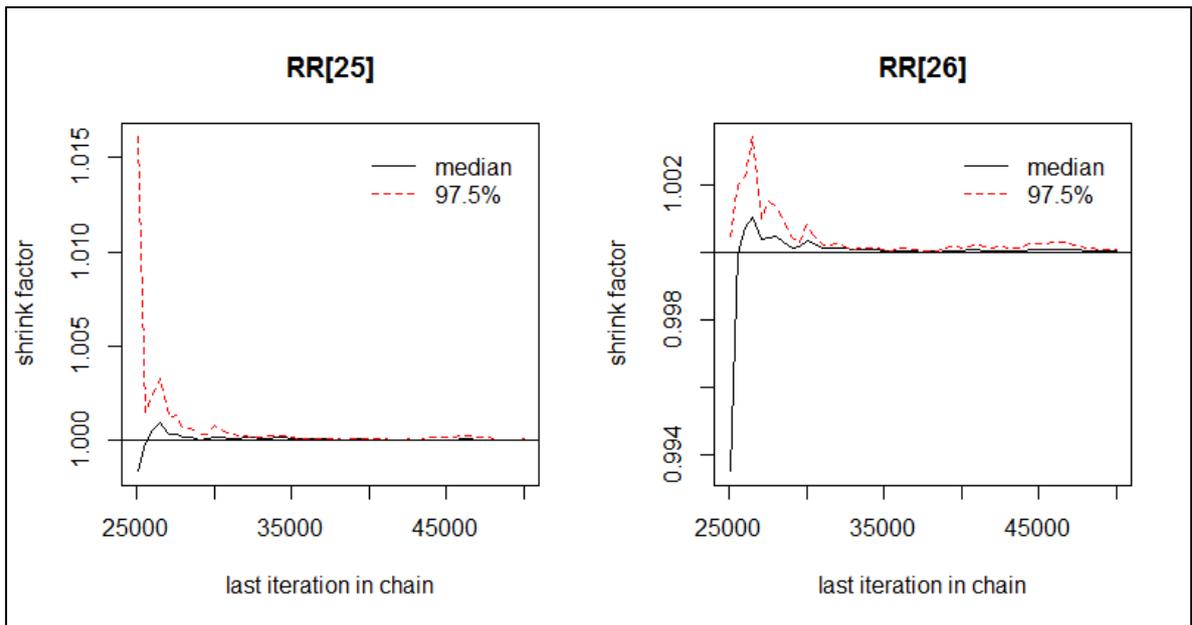
Multivariate psrf
```

**FIGURA D.7. Estadística del diagnóstico de Gelman-Rubin para los radios censales 25 y 26.**

Finalmente miramos algunos resultados estadísticos de la salida de coda (FIGURA D.9).

Se ve que la estimación puntual para el riesgo relativo es 17.18 para la región 25 y 26.77 para la región 26. En cuanto a los intervalos de credibilidad del 95%, se tuvo en cuenta, de nuevo, que se trabaja con la distribución de probabilidad y que para obtenerlos, es simplemente una cuestión de cortar los dos extremos de la distribución en 2.5% y 97.5%.

```
> gelman.plot(out.CODA[,81:82])
```



**FIGURA D.8.** Grafica del diagnóstico de Gelman-Rubin para los radios censales 25 y 26.

```

> summary(out.CODA[,81:85])

Iterations = 25001:50000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 25000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

      Mean   SD Naive SE Time-series SE
RR[25] 17.18 2.348 0.008573    0.009066
RR[26] 26.77 2.413 0.008810    0.008830

2. Quantiles for each variable:

      2.5%  25%  50%  75% 97.5%
RR[25] 12.95 15.53 17.05 18.68 22.16
RR[26] 22.26 25.09 26.69 28.35 31.73

```

**FIGURA D.9.** Estadísticos de la salida de coda para los radios censales 25 y 26.



