

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
POSTGRADO EN MODELOS ALEATORIOS

**Predicción de incidencia de malaria
mediante un modelo de retardo
distribuido autorregresivo con
componente espacial**

Autor: Lic. Francisco Laguna

Tutor Académico: Dra. María Eugenia Grillet

Asesor Académico: Dr. José R. León

**Trabajo de Grado de Maestría
Presentado ante la Ilustre
Universidad Central de Venezuela
Para optar al título de
Magister Scientiarum
Mención Modelos Aleatorios**

Caracas, 21 de Octubre de 2011

Agradecimientos

Debo agradecer a mi familia, mis profesores y todas aquellas que me han brindado su apoyo y ayuda en la realización de este trabajo.

En particular debo agradecer a los profesores José R. León y María Eugenia Grillet por su estímulo, asesoría y paciencia en la realización de todos estos estudios. Así mismo al personal del Laboratorio de Biología de Vectores por su colaboración en la obtención de los datos usado en estos estudios.

Además debo agradecer a los organismos que financiaron el estudio en cuyo marco desarrolle este trabajo de grado: Fondo Nacional de Investigaciones Científicas (FONACIT, Proyecto Misión Ciencia-UC-2008000911), y de Invensys Systems Latin America Corporation-Proyecto LOCTI 2007-2008

Resumen

Ha sido ampliamente evidenciada la influencia de la variabilidad climática en la dinámica de la incidencia de malaria en las regiones tropicales. En este trabajo se desarrolló una metodología derivada de los modelos Autorregresivos de Retardos Distribuidos (ADL), con el objetivo de extraer la influencia de las variables climáticas sobre la incidencia de malaria en el nororiente de Venezuela mediante un modelo de retardos distribuidos (DL). Al residuo del modelo de retardos distribuidos se le ajusta un modelo lineal espacio-temporal de tipo autorregresivo y los parámetros del modelo son ajustados mediante un algoritmo MCMC derivado del algoritmo de Hasting-Metropolis. El modelo más idóneo para representar la influencia de las variables climáticas fue un modelo logarítmico bivariado, usando la precipitación acumulada en combinación con El Niño como regresores, mientras que el orden del autorregresivo al que mejor se ajustó el residuo fue 2. La optimización cumplió su propósito ya que a pesar de incrementar la varianza del residuo, redujo el p-valor del test de Ljung-Box.

Índice

Agradecimientos.....	2
Resumen	3
Índice	4
1. Introducción.....	5
1.1. Breve descripción de la enfermedad.....	6
1.2. Antecedentes.....	7
1.3. Hipótesis de investigación y objetivos	8
2. Marco Teórico	10
2.1. Factores exógenos	10
2.2. Influencia espacio-temporal	11
2.3. Optimización mediante un método MCMC.	13
2.3.1. Test de pormanteau.....	15
2.3.2. Contraste de regresión.	16
2.3.3. Número de iteraciones.....	17
2.4. Esquema del método de optimización.....	18
3. Metodología.....	19
3.1. Área de estudio	19
3.2. Datos.....	20
3.3. Esquema de trabajo.....	20
4. Resultados y conclusiones	22
4.1. Trabajos preliminares	22
4.1.1. Primera fase: análisis de la densidad espectral y la ACF.	22
4.1.2. Segunda fase: identificación del Modelo ARIMAX.	28
4.1.3. Tercera fase: estimación de los parámetros del modelo ARMAX.....	31
4.2 Resultados.....	35
4.2.1. Selección de modelos	35
4.2.2. Optimización de parámetros del modelo mediante MCMC.....	39
4.3. Conclusiones.....	42
5. Bibliografía.....	43
Anexo: scripts del algoritmo de optimización.....	44

1. Introducción

La malaria se puede considerar la enfermedad parasitaria tropical más importante en el mundo¹, y la enfermedad transmisible que más muertes causa a excepción de la tuberculosis. Más de 120 millones de casos clínicos y más de 1.5 millones de muertes ocurren cada año en el mundo². A pesar de los esfuerzos hechos para reducir su transmisión ha habido poco progreso en las zonas de riesgo. De hecho, si la incidencia de la malaria continúa en su tendencia de permanente aumento, la tasa de mortalidad puede duplicarse en los próximos veinte años. Las estadísticas precisas se desconocen porque muchos casos ocurren en áreas rurales, donde el acceso a los sistemas de salud se hace difícil. A causa de esto, una buena proporción de los casos permanecen no registrados.

En Latinoamérica, la malaria constituye un serio problema de salud pública. Según la OPS, para 1999, 208 millones (44,1% de la población) vivían en zonas expuestas a algún riesgo de transmisión³. De ellos, 77 millones (16,3% de la población) viven en zonas expuestas a un riesgo moderado o elevado de transmisión. En Venezuela se ha observado en los últimos años un incremento sostenido en el registro de casos de malaria. Según el Boletín Epidemiológico N° 22 del 2010 del Ministerio de Salud no sólo el dengue tuvo un incremento sustancial en Venezuela.

Durante el 2010, se han duplicado el número de casos de

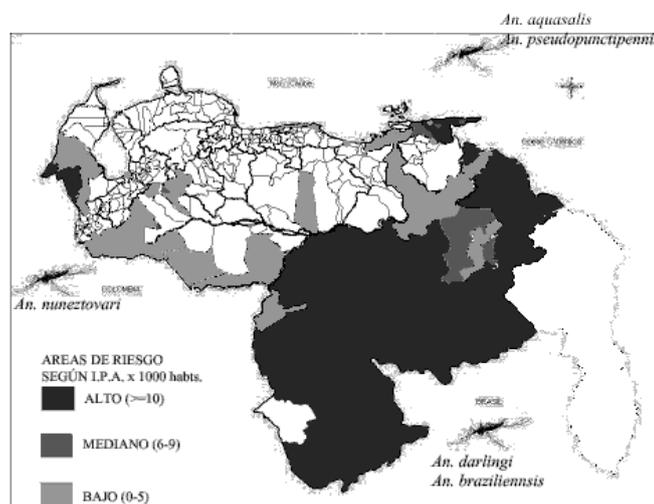


Fig. 1: Áreas de riesgo en Venezuela

Fuente: Dep. de Estadística. Div. de epidemiología. Dirección de Vigilancia Epidemiología Sanitario Ambiental.

malaria que se dieron en el mismo periodo en el 2009. Las áreas de riesgo se ubican en

¹ Who, 2009. World Malaria Report 2008.

² Breman J, *The ears of the hippopotamus*, 2001.

³ *Situación de los Programas de Malaria en las Américas*, Boletín Epidemiológico, Vol. 22 No. 1, marzo 2001

la región de Guayana (Amazonas y Bolívar), área del golfo de paria (Sucre) y el piedemonte andino (Apure, Barinas y Táchira) (Fig. 1).

1.1. Breve descripción de la enfermedad

La malaria puede ser causada por una o varias de las diferentes especies de *Plasmodium*: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale* o *Plasmodium knowlesi*. En Venezuela, los casos registrados son de *Plasmodium falciparum* y *Plasmodium vivax*. Para el presente estudio, solo se usaron los registros de *P. vivax* para la region nororiental del Estado Sucre (Figura 1). *Plasmodium vivax* es uno de los parásitos causantes de la más frecuente y extensamente distribuida forma de malaria benigna en Latinoamérica y en Venezuela, aunque no es tan virulenta o mortal como lo es *Plasmodium falciparum*, la más letal de las cuatro parasitosis.

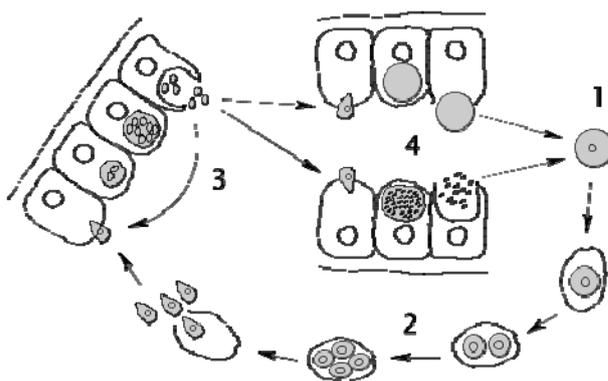


Fig. 2: Ciclo de vida del *Plasmodium*.
1-zigoto, 2-esporozoitos, 3-merozoitos, 4-gametocitos.

El ciclo de vida del *Plasmodium* consta de 4 fases básicas: zigoto, esporozoitos, merozoitos y gametocitos. Los esporozoitos se desarrollan en las glándulas salivales de la hembra de mosquitos del genero *Anopheles* y pasan al hombre durante la picadura del mosquito. Si el mosquito infectado pica a una

segunda persona sana, los esporozoitos entran a la persona a través de la saliva del mosquito y migran hacia el hígado, donde maduran y se convierten en merozoitos, los cuales penetran el torrente sanguíneo e infectan los glóbulos rojos. Los parásitos se multiplican dentro de los glóbulos que, al cabo de 48 a 72 horas, se rompen e infectan más glóbulos rojos. Los primeros síntomas se presentan por lo general de 10 días a 4 semanas después de la infección, aunque pueden aparecer incluso a los 8 días o hasta 1 año después de ésta. Luego, los síntomas de la enfermedad ocurren en ciclos de 48 a 72 horas.

La malaria esta restringida a ciertos tipos climáticos, dado que el parásito (*Plasmodium*), así como su vector (*Anopheles*), requieren climas calidos (mayor a 21 °C

para *Plasmodium*), humedad mayor a 80% para mosquitos adultos y habitat de humedales adecuados para que los mosquitos en su estado larval (o pre-adulto) completen su desarrollo.

1.2. Antecedentes

La influencia del clima sobre la incidencia de la malaria ha sido ampliamente documentada. En Baton & Ranford-Cartwright (2005) se menciona que la duración del desarrollo del *Plasmodium* en el mosquito es dependiente de la temperatura. Por ejemplo, el desarrollo de *P. falciparum* puede tomar 14 días a 26 °C pero toma 30 días a 20 °C. Esa relación entre el desarrollo del *Plasmodium* y la temperatura es mencionada también en Stressman (2010), donde además se discute sobre la importancia de la disponibilidad de agua en el desarrollo del *Anopheles*, ya que sus larvas requieren sitios donde el agua este presente por al menos 10 a 14 días. En Paaijmans et al. (2010) usando un modelo de desarrollo basado en la influencia de la temperatura sobre el desarrollo de *P. falciparum*, se muestra que las fluctuaciones de la temperatura diurna en torno a temperaturas medias superiores a 21 °C pueden acelerar el desarrollo del parásito, en cambio cuando éstas ocurren alrededor de temperaturas menores a 21° pueden desacelerar el desarrollo, comparado a lo que ocurre cuando las temperaturas son constantes. En Althizer et al (2006) se muestra como la estacionalidad puede afectar la dispersión y persistencia de enfermedades infecciosas (incluida la malaria) y que la respuesta a nivel poblacional del parasito puede variar de simples ciclos anuales a fluctuaciones mas complejas de varios años. En Basurko et al (2011), mediante análisis de correlación cruzada se comprueba la influencia de las variables meteorológicas y el nivel del río sobre la incidencia de malaria. En Colón-González (2011) se hace uso de modelos de regresión lineal múltiple para estudiar la relación entre la variabilidad climática y la incidencia de dengue. En Sáez-Sáez (2007) se estudia la influencia de la precipitación y la temperatura sobre la incidencia de malaria y variables climáticas en el estado Sucre, Venezuela, en donde se recomienda que dada la dinámica de la transmisión de la malaria, los análisis de la relación entre las variables climáticas y la incidencia de la enfermedad deben examinarse a escalas de tiempo menores que un mes.

En la última década se ha dado mayor importancia al factor espacial en el estudio de la incidencia de la malaria. En Barrera et al (1998) se hace uso de Sistemas de Información Geográfica (SIG) y Sensores remotos para hacer un análisis espacio-

temporal de la dinámica de malaria en el estado Sucre. En Briët (2008) se reporta una correlación espacial significativa entre la estacionalidad de la precipitación y la estacionalidad de la malaria en un estudio hecho en Sri Lanka. En Parham (2010) se plantea que un modelo más realista de la transmisión de la malaria necesita la selección de una escala más apropiada para desarrollar una mejor predicción. Varios autores más han propuesto modelos espacio-temporales de la incidencia de malaria tales como Barrera et al (1999), Rogers J et al (2002), Hoshen & Morse (2004), Sáez-Sáez (2007), Chowell et al (2009), Grasso et al (2010) y Grillet et al (2010).

1.3. Hipótesis de investigación y objetivos

Como ya ha sido referido en los antecedentes la incorporación del clima se hace de vital importancia en la modelización de la malaria. Asimismo incorporación del factor espacial suministra un enfoque más realista de la dinámica de la enfermedad ya que esta no ocurre en un entorno homogéneo, si no muy diverso en el espacio y el tiempo. De modo tal que la pregunta central de este trabajo de grado sería. ¿La dinámica de la malaria en una localidad puede ser explicada a partir de la dinámica de variables meteorológicas como precipitación y temperatura (influencias exógenas) así como por la dinámica de la malaria en otras localidades (influencia espacio-temporal)? En esta pregunta se pueden extraer las dos hipótesis de investigación de este trabajo de grado:

- La dinámica de la malaria puede ser explicada a partir de la dinámica de variables meteorológicas.
- La dinámica de la malaria en cualquier localidad puede ser explicada a partir de la estructura espacio-temporal de la dinámica regional de la malaria.

El objetivo de esta tesis es plantear un modelo que puede verificar ambas hipótesis. El modelo que planteo para verificar esta hipótesis estaría derivado de un modelo ADL o autorregresivo con retardos distribuidos, el cual pertenece a la familia de modelos dinámicos lineales, siendo una de sus propiedades más destacables que los estimadores de los coeficientes del modelo son insesgados, es decir que el estimador se corresponde a su esperanza. El modelo ADL ofrece la ventaja de poder separar la influencia exógena, la cual sería manejada por el componente de retardos distribuidos, de la influencia espacio-temporal, la cual sería tratada mediante una combinación lineal de autorregresivos de las distintas localidades. Este modelo será tratado con mas detalle en la sección correspondiente al marco teórico.

Algo que debe ser tomado en cuenta es el comportamiento del residuo o perturbaciones (z_i), ya que según el teorema de Gauss-Markov, al estimar un modelo de regresión lineal, si las perturbaciones son no correlacionadas, idénticamente distribuidas e independientes de las variables explicativas, entonces los estimadores mínimos cuadrados son óptimos o de mínima varianza dentro de la clase de los estimadores centrados que son funciones lineales de las observaciones. Se hace necesario verificar que el residuo cumpla las hipótesis:

- El residuo z se distribuye como un ruido blanco, es decir que es no correlacionado e idénticamente distribuido.
- El residuo z es ortogonal a las variables explicativas.

¿En caso de que esto no ocurra los parámetros del modelo pueden ser ajustados de modo que el residuo pueda cumplir estos requisitos? Para ajustar los parámetros del modelo de modo que cumpla ambas hipótesis se implementó un método derivado del algoritmo de Hasting-Metropolis.

Con la metodología que se presenta se busca:

- Establecer un modelo de regresión que relacione la dinámica de la incidencia de malaria con variables meteorológicas y climáticas (variables exógenas)..
- A partir del residuo estimar un modelo espacio-temporal lineal, que explique como se relaciona la dinámica de una localidad con la de las demás localidades.
- Ajustar mediante un método MCMC los parámetros del modelo de modo que cumpla con los supuestos necesarios de un modelo lineal.

2. Marco Teórico

El procedimiento inicial consistió en 3 fases

- Se estimó la influencia de las variables exógenas sobre la incidencia de malaria mediante un modelo de regresión lineal múltiple.
- Se estimó la influencia temporal (es decir la influencia de las observaciones previas) mediante un modelo ARMA(p, q).
- Se estimó la influencia espacial (es decir, la influencia de las observaciones de otras localidades) mediante un modelo de regresión lineal múltiple.

Posteriormente los parámetros espacio-temporales (fase 2 y 3) serán ajustados mediante un algoritmo de Hasting-Metropolis modificado.

2.1. Factores exógenos

La influencia de las variables exógenas (tales como variables climáticas y otros factores que pudiesen influir en la dinámica de la enfermedad) sobre la incidencia parasitaria $\{y_t\}$ se expresa mediante un modelo de retardos distribuidos RD(r_1, \dots, r_b). Su expresión estaría dada por

$$\mu_t = \beta^* + \sum_{i=1}^b \sum_{j=0}^{r_i} \beta_{i,j} x_{i,t-j}$$

siendo $\{x_{i,t}\}$ la i -ésima variable exógena del modelo, y r_i el número máximo de retardos para $\{x_{i,t}\}$. Los coeficientes $\beta_{i,j}$ y β^* se estiman mediante una regresión múltiple. La bondad de ajuste para cada modelo RD(r_1, \dots, r_b), usado se medirá mediante su coeficiente de determinación corregido por el número de grados de libertad \bar{R}^2 , ya que éste tiene la ventaja sobre el estadístico R^2 en que penaliza el número de variables regresoras y por lo tanto, los r_k muy grandes. Éste estaría definido como

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n - (k+1)}$$

De modo tal que $\bar{R}^2 \leq R^2$.

De este modo, la incidencia de malaria (y_t) queda expresado como una combinación lineal de las variables exógenas ($x_{j,t}$), más un residuo (w_t). Se calcula la serie $\{w_t\}$ restando $\{\mu_t\}$ a $\{y_t\}$.

$$\begin{aligned} y_t &= \mu_t + w_t \\ y_t &= \beta^* + \sum_{i=1}^b \sum_{j=0}^{r_k} \beta_{i,j} x_{i,t-j} + w_t \end{aligned}$$

2.2. Influencia espacio-temporal

La influencia temporal se representó ajustando a la serie $\{w_t\}$ un modelo AR(p,q). De tal forma que $\{y_t\}$ se ajustó a un modelo autorregresivo de retardos distribuidos (ADL) de la forma

$$y_t = \sum_{i=1}^p \phi_i w_{t-i} + \left(\beta^* + \sum_{i=1}^b \sum_{j=0}^{r_k} \beta_{i,j} x_{i,t-j} \right) + e_t$$

donde ϕ_1, \dots, ϕ_p son los coeficientes del modelo autorregresivo y e_t es un ruido blanco de varianza σ_e^2 . Aunque en el trabajo previo al residuo se le ajustó un modelo ARMA, para la metodología que se sigue en este trabajo se prefiere ajustar a un modelo autorregresivo. Si bien con un ARMA se podría obtener un mejor ajuste, la función de verosimilitud se hace más difícil de desarrollar.

La influencia espacial se incorpora cambiando la formulación clásica del modelo autorregresivo por la siguiente

$$w_t = \sum_{i=1}^p \phi_{k,i} w'_{k,t-i} + z_{k,t}$$

siendo

$$w'_{k,t} = \lambda_{k,k} w_{k,t} + \sum_{l \in L_k, l \neq k} \lambda_{k,l} y_{l,t}$$

donde $\lambda_{k,l}$ son los pesos que tiene las series $\{y_{l,t}\}$ y la misma $\{w_{k,t}\}$ sobre la dinámica de $\{w_{k,t}\}$, mientras ϕ_1, \dots, ϕ_p son los coeficientes del modelo autorregresivo estimados anteriormente. De modo que

$$w_{k,t} = \sum_{j=1}^p \phi_{k,j} \left(\lambda_{k,k} w_{k,t-j} + \sum_{l \in L_k, l \neq k} \lambda_{k,l} y_{l,t-j} \right) + z_{k,t}$$

Esto puede ser reescrito de tal modo que

$$w_{k,t} = \lambda_{k,k} \phi'(B) w_{k,t} + \sum_{l \in L_k} \lambda_{k,l} \phi'(B) y_{l,t} + z_{k,t}$$

siendo $\phi'(B) = \phi_1 z + \dots + \phi_p z^p$.

Para estimar los coeficientes $\lambda_{k,l}$ primero escribamos la expresión anterior de forma matricial, quedando

$$\vec{w}_t = V \times \vec{\lambda} \quad (*)$$

siendo $\vec{w}_t = (w_{p+1}, \dots, w_{t+1})$, $\lambda = (\lambda_1, \dots, \lambda_{k-1}, \lambda_k, \lambda_{k+1}, \dots, \lambda_{|L_k|})$ y

$$V = \begin{bmatrix} \phi'(B) y_{1,t} & \cdots & \phi'(B) y_{k-1,t} & \phi'(B) w_{k,t} & \phi'(B) y_{k+1,t} & \cdots & \phi'(B) y_{|L_k|,t} \\ \phi'(B) y_{1,t-1} & \cdots & \phi'(B) y_{k-1,t-1} & \phi'(B) w_{k,t-1} & \phi'(B) y_{k+1,t-1} & \cdots & \phi'(B) y_{|L_k|,t-1} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \phi'(B) y_{1,p} & \cdots & \phi'(B) y_{k-1,p+1} & \phi'(B) w_{k,p+1} & \phi'(B) y_{k+1,p+1} & \cdots & \phi'(B) y_{|L_k|,p+1} \end{bmatrix}$$

Pre-multiplicando a ambos lados de la expresión (*) por V^{-1} , obtenemos

$$\vec{\lambda} = V^{-1} \vec{w}_{t+1}$$

Una vez estimado $\vec{\lambda}$, se puede calcular el predictor $p \vec{w}_t$, que esta definido como

$$p \vec{w}_{t+1} = V \times \vec{\lambda}$$

Al restar $p \vec{w}_{t+1}$ a \vec{w}_t se obtiene el residuo $\{z_{k,t}\}$ que debe cumplir las siguientes condiciones

- $\{z_{k,t}\} \sim \text{WN}(0, \sigma_k^2)$
- $z_{k,t} \perp y_{l,t} \forall k, l \in L_k$ y $z_{k,t} \perp w_{k,t}$

De este modo $\vec{w}_k \sim N(\vec{\omega}_k, \Gamma_k)$, donde $\Gamma_k = I \cdot \sigma_k^2$ y $\vec{\omega}_k = V \times \vec{\lambda}$.

Así la función de verosimilitud estaría dada por

$$L(\vec{w}_k | \sigma_k^2, \Theta_k) = \frac{1}{(2\pi)^{n/2} \sigma_k^n} \exp\left(-\frac{1}{2\sigma_k^2} \|\vec{w}_k - \vec{\omega}_k\|^2\right)$$

donde Θ_k es un vector formado por los parámetros del modelo (ϕ_k y λ_k)

2.3. Optimización mediante un método MCMC.

Al estimar por separado los coeficientes autorregresivos, ϕ_1, \dots, ϕ_p y los coeficientes espaciales $\lambda_{k,l}, k, l \in L_k$ surge el problema de que el residuo no necesariamente cumple las condiciones mencionadas anteriormente, es decir que se distribuya como un ruido blanco y que sea ortogonal al residuo $w_{k,t}$ y las variables exógenas en las localidades vecinas. Un modo de solventar este problema es usar un algoritmo de Metropolis-Hasting modificado. El algoritmo de Metropolis-Hastings es un método MCMC (Markov chain Monte Carlo) para la obtención de una secuencia de muestras aleatorias de una distribución de probabilidad cuyo muestreo directo es difícil. Esta secuencia se puede utilizar para aproximar la distribución (es decir, para generar un histograma) o para calcular una integral (como un valor esperado).

El algoritmo genera una cadena de Markov en el que cada estado Θ_{m+1} depende solo del estado previo Θ_m . El algoritmo usa un kernel $Q(\Theta_m, \Theta')$, el cual depende únicamente del estado Θ_m , para generar una nueva muestra Θ' . El kernel de transición será una función gaussiana centrada en el actual estado Θ_m .

$$Q(\Theta_m, \Theta') \sim N(\Theta_m, \bar{\sigma}_Q^2 I)$$

La muestra propuesta (Θ') es aceptada como el siguiente valor ($\Theta_{m+1} = \Theta'$) si un valor v tomado de una uniforme (0,1) satisface que $v \leq \alpha(\Theta_m, \Theta')$. Si no es aceptada, entonces el actual valor Θ es retenido: $\Theta_{m+1} = \Theta_m$

La función $\alpha(\Theta_m, \Theta')$ está definida del siguiente modo

$$\alpha(\Theta_m, \Theta') = \min \left\{ \frac{L(\vec{w}_k | \Theta') Q(\Theta', \Theta_m)}{L(\vec{w}_k | \Theta_m) Q(\Theta_m, \Theta')}, 1 \right\}$$

Pero al ser $Q(\Theta_n, \Theta_{n+1})$ una función gaussiana ella es simétrica y entonces

$$\begin{aligned} \alpha(\Theta_m, \Theta') &= \min \left\{ \frac{L(\vec{w}_k | \Theta')}{L(\vec{w}_k | \Theta_m)}, 1 \right\} \\ &= \min \left\{ \left(\frac{\sigma_m}{\sigma'} \right)^n \frac{\exp \left[-\frac{1}{2\sigma'^{2n}} \|\vec{w}_k - \vec{w}'_k\|^2 \right]}{\exp \left[-\frac{1}{2\sigma_m^{2n}} \|\vec{w}_k - \vec{w}_k\|^2 \right]}, 1 \right\} \end{aligned}$$

Dado que σ^2 varía en función de los cambios de θ_k y λ_k , entonces se podría cambiar σ^2 por $\hat{\sigma}^2$, de tal modo que

$$\begin{aligned} \alpha(\Theta_m, \Theta') &= \min \left\{ \left(\frac{\hat{\sigma}_m}{\hat{\sigma}'} \right)^n \frac{\exp \left[-\frac{1}{2\hat{\sigma}'^{2n}} \|\vec{w}_k - \vec{w}'_k\|^2 \right]}{\exp \left[-\frac{1}{2\hat{\sigma}_m^{2n}} \|\vec{w}_k - \vec{w}_k\|^2 \right]}, 1 \right\} \\ &= \min \left\{ \left(\frac{\hat{\sigma}_m}{\hat{\sigma}'} \right)^n, 1 \right\} \end{aligned}$$

Resulta que dado que con un n grande el cociente de las varianzas tiende muy rápidamente a cero, cualquier nueva muestra sería rechazada. Para solucionar esto, en lugar de usar n como exponente del cociente se usa una constante φ , de modo que la función $\alpha(\Theta_m, \Theta')$ queda definida así:

$$\alpha(\Theta_m, \Theta') = \min \left\{ \left(\frac{\hat{\sigma}_m}{\hat{\sigma}'} \right)^\varphi, 1 \right\}$$

Para asegurar que $\{z_t\}_k$ cumple con las condiciones de que se distribuya como un ruido blanco y que sea ortogonal a las series $\{y_t\}_k$ de las localidades vecinas se debe modificar el algoritmo para truncar la distribución de Θ de modo que esta cumpla tales condiciones. Antes de generarse ν y verificarse si es menor que $\alpha(\Theta_m, \Theta')$, se deben realizar 2 pasos adicionales:

- Verificar que Θ' genera un residuo que se distribuya como ruido blanco (test de pormanteau)
- Verificar que Θ' genera un residuo ortogonal a las series $\{y_t\}_k$ de las localidades vecinas (contraste de regresión).

2.3.1. Test de pormanteau

Un test de pormanteau está diseñado para verificar si un conjunto de autocorrelaciones de una serie de tiempo es diferente de cero, lo cual implica bajo la hipótesis de gaussianidad que las variables no son iid gaussianas. Se usara el test de Ljung-Box, uno de los test de pormanteau mas conocidos. El test esta definido de la siguiente manera.

- H_0 : Los datos son iid.
- H_a : Los datos no son iid

El estadístico del test esta definido como

$$Q_{LB} = n(n+2) \sum_{j=1}^h \frac{\hat{\rho}^2(j)}{n-j}$$

donde n es el tamaño de la muestra, $\hat{\rho}^2(j)$ es la autocorrelación muestral a j rezagos y h es el numero de rezagos a ser comprobados. Q_{LB} se distribuye como una χ^2 con h grados de libertad

En la inicialización del algoritmo de Metropolis-Hasting se realiza un test de Ljung-Box y se guarda su p-valor. Luego para cada Θ' generado se realiza un test de Ljung-Box y se verifica si es mayor que un $\alpha_{LB} \in (0,1)$ asignado arbitrariamente. Si es mayor que este se pasa a verificar la ortogonalidad. De lo contrario, entonces se verifica si su p-valor es mayor que el p-valor del ultimo Θ' aceptado. Si se verifica esto se pasa a verificar la ortogonalidad. De lo contrario se rechaza Θ' y $\Theta_{m+1} = \Theta_m$.

2.3.2. Contraste de regresión.

Un modo de verificar la ortogonalidad del nuevo residuo $\{z'_{k,t}\}$ respecto a las series $\{y_{k,t}\}$ de las localidades vecinas es realizar una regresión múltiple, tomando como variable respuesta a $\{z'_{k,t}\}$ y como variables regresoras a las series $\{y_{l,t}\}$ de las localidades vecinas y de la serie $\{y_{k,t}\}$ y a esa regresión se le aplica un test de contraste de regresión. El test de contraste de regresión se basa en descomponer la varianza de la variable respuesta en la varianza explicada por el modelo de regresión más la varianza no explicada y residual, permitiendo contrastar si el modelo es significativo o no. Bajo este razonamiento se puede hacer la siguiente descomposición:

$$\sum_{i=1}^{n-p} (z_i - \bar{z})^2 = \sum_{i=1}^{n-p} (z_i - \hat{z}_i)^2 + \sum_{i=1}^{n-p} (\hat{z}_i - \bar{z})^2$$

De esta igualdad se construye la siguiente tabla ANOVA

Tabla ANOVA del modelo de regresión múltiple			
Fuente de variación	Suma de cuadrados	Grados de libertad	Varianzas
Modelo de regresión	$scE = \sum_{i=1}^n (z_i - \hat{z}_i)^2$	$ L_k $	$\hat{s}_e^2 = \frac{scE}{ L_k }$
Residual	$scR = \sum_{i=1}^n (\hat{z}_i - \bar{z})^2$	$n - (L_k + p - 1)$	$\hat{s}_r^2 = \frac{scR}{n - (L_k + p - 1)}$
Global	$scG = \sum_{i=1}^n (z_i - \bar{z})^2$	$n - (p + 1)$	$\hat{s}_z^2 = \frac{scG}{n - (p + 1)}$

De esta tabla ANOVA se deduce el siguiente contraste acerca de la influencia “conjunta” del modelo de regresión en la variable respuesta.

- $H_0 : b_i = 0 \forall i \in L_k$
- $H_1 : \text{al menos un } b_i \neq 0.$

Si H_0 es cierto ninguna de las variables regresoras influye en la variable respuesta (es decir que el modelo no explica la dinámica de la variable respuesta). En este supuesto se verifica que $\hat{y} \approx \bar{y} \Rightarrow scE \approx 0$. Por ser ésta una medida absoluta se compara con la varianza residual, lo que lleva a utilizar como estadístico del contraste el siguiente

$$\hat{F} = \frac{\hat{s}_e^2}{\hat{s}_r^2}.$$

Bajo la hipótesis nula y por la hipótesis de independencia se sigue que \hat{F} sigue una distribución F (Contraste de la F) con $|L_k|$ y $n - (|L_k| + p - 1)$ grados de libertad,

$$\hat{F} | H_o = \frac{\hat{S}_e^2}{\hat{S}_r^2} \sim F_{|L_k|, n - (|L_k| + p - 1)}.$$

En la inicialización se aplica el test de contraste de regresión al residuo $\{z'_i\}_k$ y se guarda el p-valor del estadístico \hat{F} . Luego para cada Θ' generado se realiza un test de contraste de regresión y se verifica si es mayor que un $\alpha_F \in (0,1)$ asignado arbitrariamente. Si es mayor que éste se realiza la verificación clásica del algoritmo de Metropolis-Hasting. De lo contrario, entonces se verifica si su p-valor es mayor que el p-valor del último Θ' aceptado. Si se verifica esto se realiza la verificación clásica. De lo contrario se rechaza Θ' y $\Theta_{m+1} = \Theta_m$.

2.3.3. Número de iteraciones

Debe ser tomado en cuenta que el algoritmo de Metropolis-Hasting lo que hace es construir una cadena de Markov para así simular la distribución de la variable de interés. De modo que para asegurar que la cadena ha alcanzado su estado estacionario se descartan las primeras n_b iteraciones. Mientras mayor sea n_b es más probable que las iteraciones restantes $(M - n_b)$ pertenezcan a una cadena estacionaria.

2.4. Esquema del método de optimización

Inicialización

1. se estima $\{\mu_t\}_k$.
2. se ajusta al residuo $\{w_{k,t}\}$ un modelo AR(p) y se guardan los coeficientes $\phi_{k,j}$.
3. se construye la matriz V a partir de las series vecinas $\{y_{l,t}\}$, $l \neq k$ y el residuo $\{w_{k,t}\}$ y con ella se estima los coeficientes λ_k .
4. se calcula la varianza del residuo $\{z_{k,t}\}$ y se guarda como V_{opt} .
5. se aplica a $\{z_t\}_k$ un test de Ljung-Box y el p-valor del test se guarda como LB_{opt} .
6. se realiza una regresión múltiple $\{z_t\}_k$ con respecto a $\{y_t\}_l$ para todo $l \neq k$. El p-valor del estadístico F se guarda como F_{opt} .

Muestreo mediante MCMC

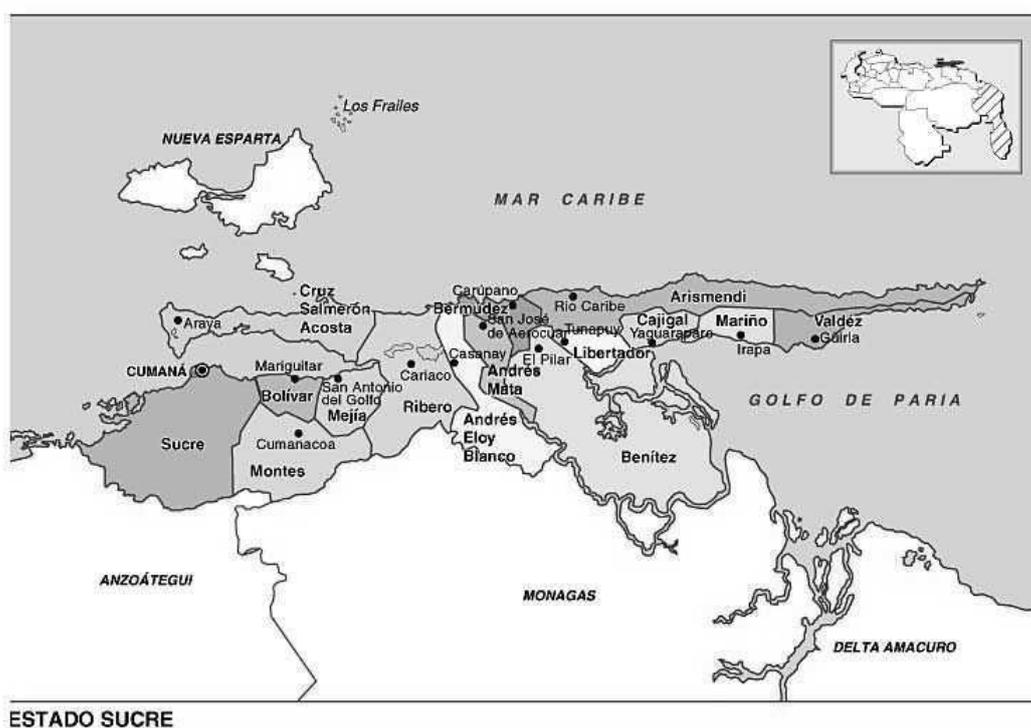
1. Dado Θ_m se genera Θ_{ult} , mediante $Q(\Theta_m, \Theta')$.
- 2.a se aplica un test de Ljung-Box a $\{z'_t\}$ y si su p-valor (LB_{ult}) es mayor que LB_{umbral} (escalar arbitrario): se pasa a 3.a.
si no: a 2.b
- 2.b si LB_{ult} es mayor que LB_{opt} : $LB_{\text{opt}} = LB_{\text{ult}}$ se pasa a 3.a.
Si no: $\Theta_{m+1} = \Theta_m$ y $m = m+1$ y se vuelve a 1.
- 3.a se realiza contraste de regresión a $\{z'_t\}$ y si el p-valor del estadístico F (F_{ult}) es mayor que F_{umbral} (escalar arbitrario): se pasa a 4.
Si no: se pasa a 3.b.
- 3.b si F_{ult} es mayor que F_{opt} : $F_{\text{opt}} = F_{\text{ult}}$ se pasa a 4.
Si no: $\Theta_{m+1} = \Theta_m$ y $m = m+1$ y se vuelve a 1.
4. se genera u a partir de $U(0,1)$.
Si $u \leq \alpha(\Theta_{\text{opt}}, \Theta_{\text{ult}}) = \min\left\{\left(V_{\text{opt}}/V_{\text{ult}}\right)^n, 1\right\}$:
 $\Theta_{\text{opt}} = \Theta_{\text{ult}}$, $V_{\text{opt}} = V_{\text{ult}}$, $m = m+1$ y se vuelve a 1.
Si no: $\Theta_{m+1} = \Theta_m$, $m = m+1$ y se vuelve a 1.

Este procedimiento se repite m veces. A partir de las últimas $(M - n_b)$ muestras se estima la media muestral $\bar{\Theta}$. n_b es el número de iteraciones que requieren ser quemadas, es decir descartadas para asegurar que las iteraciones usadas en el cálculo de la media pertenecen a una cadena estacionaria.

3. Metodología

3.1. Área de estudio

Sucre es uno de los 23 estados de Venezuela, ubicado en la región Nororiental del país (Fig 1), limita al norte con el Mar Caribe, al sur por los Estados Anzoátegui y Monagas, por el este con el golfo de Paria y por el oeste con el golfo de Cariaco. El estado está compuesto por 15 municipios divididos en 54 parroquias. En la zona del litoral Occidental se observa un tipo de clima semiárido, registrándose en Cumaná 24-26 °C de temperatura media anual con una pluviosidad de 375 mm. En la faja paralela meridional se observa transicionalmente un clima tropical lluvioso de sabana, que se extiende hasta la zona de la vertiente litoral al mar Caribe. En el Golfo de Paria la media de lluviosidad anual aumenta entre 1.200 a más de 2.000 mm. con un clima boscoso decidido. En la zona de la Serranía del Interior se registra una sección en donde prevalece el clima tropical de altura.



3.2. Datos

Como variable respuesta se trabajó con la incidencia parasitaria semanal (IPS) de malaria de los últimos 8 años, estratificada dicha data por parroquias. El IPS es definido como el número de casos nuevos por semana estandarizados por el número poblacional de la entidad.

Como variables exógenas se usaron:

- La precipitación y temperatura media semanal registrada en la estación meteorológica de la aviación de Guiria, municipio Valdez.
- La temperatura de la superficie del mar (El Niño⁴) registrada en el área ubicada entre los paralelos 5° N y 5°S y los meridianos 170°W y 120°W, como una alternativa a la temperatura local (registrada en Guiria).

3.3. Esquema de trabajo

Para determinar para cual valor de r_i se obtiene el mejor ajuste, se realizó una regresión múltiple con distintos $r \in \{4, 8, \dots, 76, 80\}$, midiendo el ajuste mediante su R^2 ajustado, de modo de penalizar valores altos de r . Así mismo se probó la lluvia acumulada con distintos valores de $h \in \{1, 2, \dots, 25, 26\}$. Se evaluaron los siguientes modelos de regresión:

Tabla 1: Modelos evaluados

Univariado	No logarítmico	Temperatura media semanal	
		Humedad media semanal	
		Precipitación Acumulada	
logarítmico		Temperatura media semanal	
		Humedad media semanal	
		Precipitación Acumulada	
Bivariado	No logarítmico	Precipitación Acumulada + Temperatura media semanal	
		Precipitación Acumulada + Humedad media semanal	
	logarítmico		Precipitación Acumulada + Temperatura media semanal
			Precipitación Acumulada + Humedad media semanal

Al aplicar la transformada logarítmica tanto al IPS como a la precipitación acumulada, en lugar de aplicar $\log(x)$ se aplico $\log(x + 1)$, para solucionar el problema de los ceros presentes en la serie. Para la selección de modelos se usaron únicamente los registros de malaria de los municipios Benítez y Cagigal, los cuales presentan una alta y continua

⁴ (SST 3+4) (<http://www.cpc.ncep.noaa.gov/data/indices/>)

incidencia de malaria durante el periodo de estudio. Una vez elegido la combinación optima de parámetros para cada modelo, los distintos modelos se compararon entre si en base a su valor BIC y el p-valor de su test de Ljung-Box.

Posteriormente, se probó el modelo seleccionado, usando distintas medidas de temperatura. Además de la temperatura media semanal se usó la temperatura máxima semanal y la minima semanal y en adición se utilizaron los datos del Niño. Para esta fase del estudio se utilizaron los municipios A. E. Blanco, A. Mata, Arismendi, Benítez, Bermúdez, Bolívar, C. S. Acosta, Cajigal, Libertador, Mariño, Mejía, Montes, Ribero, Sucre y Valdez. Los resultados se compararon en base a su coeficiente de determinación, R^2 .

Una vez seleccionado el modelo óptimo, se construyó con los registros de los municipios A. E. Blanco, A. Mata, Arismendi, Benítez, Bermúdez, Cajigal, Libertador, Mariño, Ribero y Valdez del año 2001 al 2008, la rejilla sobre la que se iba a correr el modelo. Esta rejilla se estratificó por parroquia y se excluyeron aquellas parroquias en las que no hubo en ningún momento registros de casos de malaria. Se compararon los siguientes estadísticos antes y después de aplicarse la optimización: Varianza del residuo z , p-valor del test de Ljung-Box y p-valor del test F. Esto se hizo para verificar la eficacia de la optimización.

4. Resultados y conclusiones

4.1. Trabajos preliminares

En esta sección hablaré de los resultados obtenidos durante mi pasantía realizada en el laboratorio de vectores, ya que en base a estos resultados planteo el desarrollo que constituye el actual trabajo de grado. Para estos estudios se usaron los registros de incidencia de malaria de los municipios Benítez y Cagigal del estado Sucre del año 2001 al 2008, de todos los municipios del estado Amazonas del año 1998 al 2009 y de todos los municipios del estado Bolívar del año 1996 al 2008. Como variables regresoras se usaron los registros de precipitación, temperatura y humedad obtenidos en las estaciones meteorológicas de la aviación en Guiria, estado Sucre y Tumeremo, estado Bolívar del año 1997 al 2008. Para el estado Amazonas se usó el registro de precipitación de la estación meteorológica de Pto. Ayacucho del año 1997 al 2008 y el nivel de río del Orinoco a la altura de Pto. Ayacucho del año 1997 al 2009.

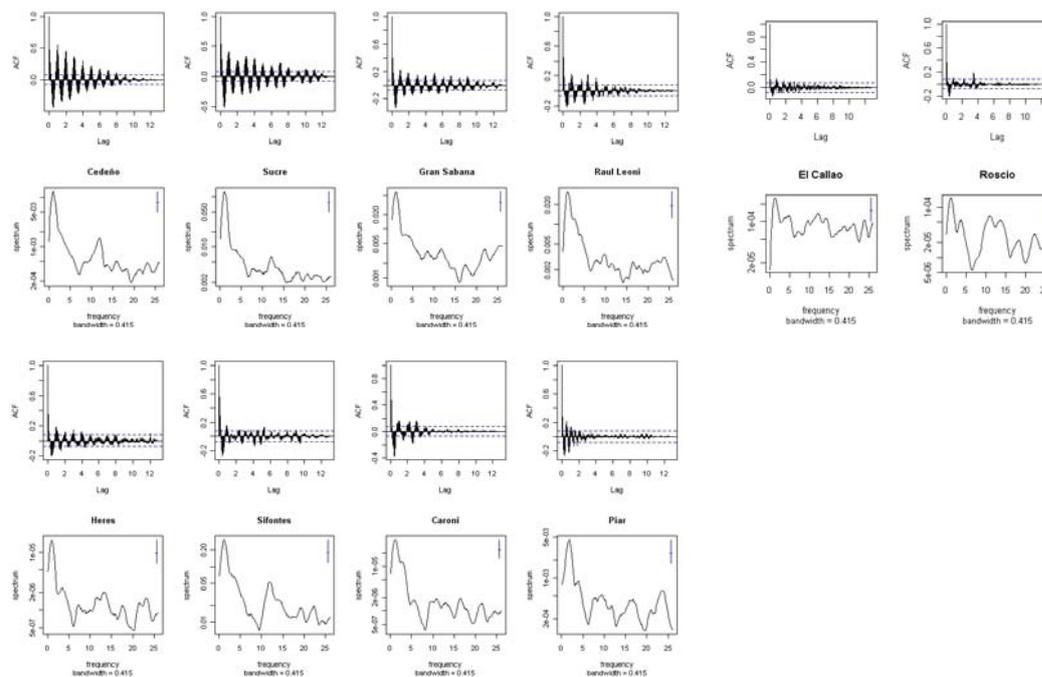
4.1.1. Primera fase: análisis de la densidad espectral y la ACF.

En la primera fase se estimó la función de autocovarianza (ACF) y la densidad espectral para cada municipio (ver Fig. 3). Luego se compararon estas buscando similitudes entre ellas de modo que se pudiesen proponer regiones con dinámicas homogéneas. En el estado Bolívar se extrajo la tendencia a cada serie para que ésta no distorsionase el componente estacional.

En Bolívar se observó un patrón estacional muy nítido en los municipios de Cedeño, Sucre, Gran Sabana y Raúl Leoni. Junto al ciclo anual se observó otro en una frecuencia entre 10 y 12 ciclos por años (un periodo de 4 a 5 semanas)

Bolívar

Fig. 3: ACF y densidad espectral por municipio en el estado Bolívar



El patrón estacional no se ve tan nítido en los municipios Heres, Sifontes, Piar y Caroní, Roscio y Callao. Se puede detectar la existencia del ciclo de 4 a 5 semanas ya mencionado, salvo en Caroní, Piar y el Callao que no predomina ningún otro ciclo aparte del anual

Es interesante haber detectado este ciclo circa mensual en Tabla 2

	% ceros	IPA
Cedeño	0.0385	0,32
Sucre	0.0343	0,97
Raúl Leoni	0.0330	0,70
Gran Sabana	0.0536	0,66
Sifontes	0.0288	3,13
Piar	0.1126	0,25
Caroní	0.3104	0,01
El Callao	0.6305	0,04
Roscio	0.8791	0,01

como se ve en la tabla 2), de modo que no se puede lograr una buena estimación.

Se usaron correlaciones cruzadas para detectar similitud entre series. Por un lado se usaron las series crudas para comparar las tendencias y por otro se usaron las series con la tendencia extraída mediante medias móviles para comparar el componente estacional. Se asume que mientras más simétrica sea la correlación cruzada mas similares son las series.

Se puede ver que los municipios Caroní, Piar, Heres y El Callao poseen una evolución bastante similar. Por otro lado este conjunto de municipios guardan una correlación semejante con Sifontes y con Heres. Aunque de las correlaciones cruzadas entre las series sin tendencia no se puede sacar nada en claro. En realidad estos municipios mostraron una estructura espectral más bien abigarrada. En cambio con los municipios Cedeño, Sucre, Raúl Leoni y Gran Sabana se encontró una gran similitud a nivel de su estructura estacional y hasta cierto punto en su evolución en particular entre Cedeño y Sucre

Fig. 4: Correlaciones Cruzadas de los municipios Heres, Sifontes, Caroní, Piar, El Callao y Roscio.

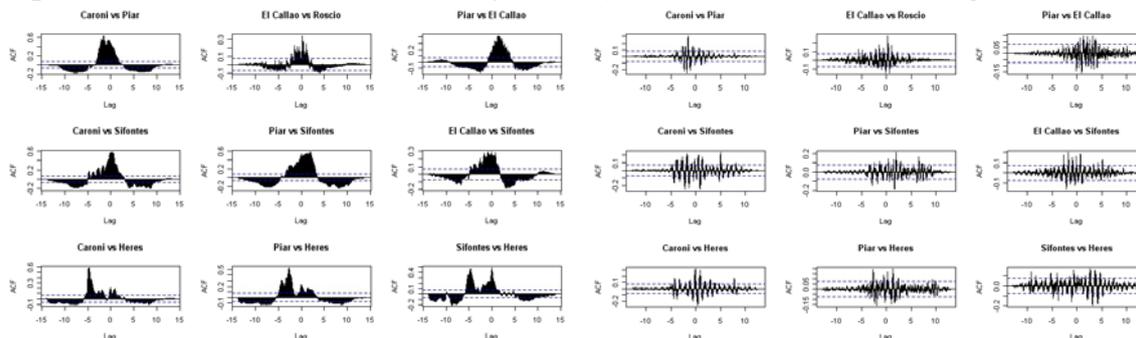
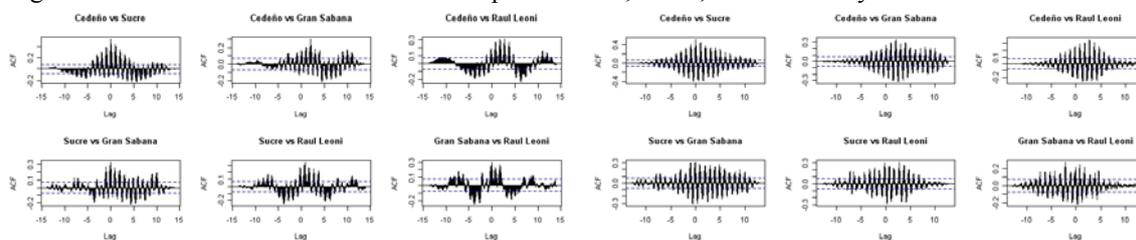
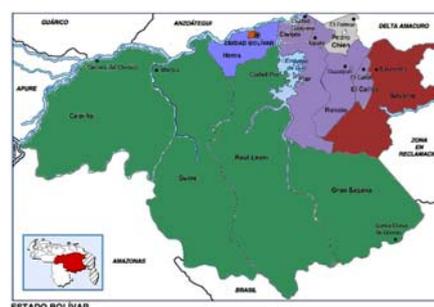


Fig. 5: Correlaciones Cruzadas de los municipios Cedeño, Sucre, Gran Sabana y Raúl Leoni



En base a esto considero que el modo más adecuado de agrupar los municipios de Bolívar sería:

- Cedeño-Sucre-Raúl Leoni-Gran Sabana (CSRG)
- Piar-Caroní- El Callao-Roscio (PCCR)
- Sifontes
- Heres



El sector CSRG es el que muestra la estructura estacional más nítida, mientras PCCR no muestra una estructura estacional clara. Sifontes y Heres están en una situación intermedia a estos dos. En general PCCR, CSRG y Sifontes son muy similares en su evolución, mas Heres difiere de todos ellos. A nivel de la estructura estacional, CSRG,

Sifontes y Heres guardan cierta similitud entre ellos en cambio, como era de esperarse por su estructura estacional tan irregular no guarda similitud con ningún otro sector.

Fig. 6: ACF y densidad espectral por sector

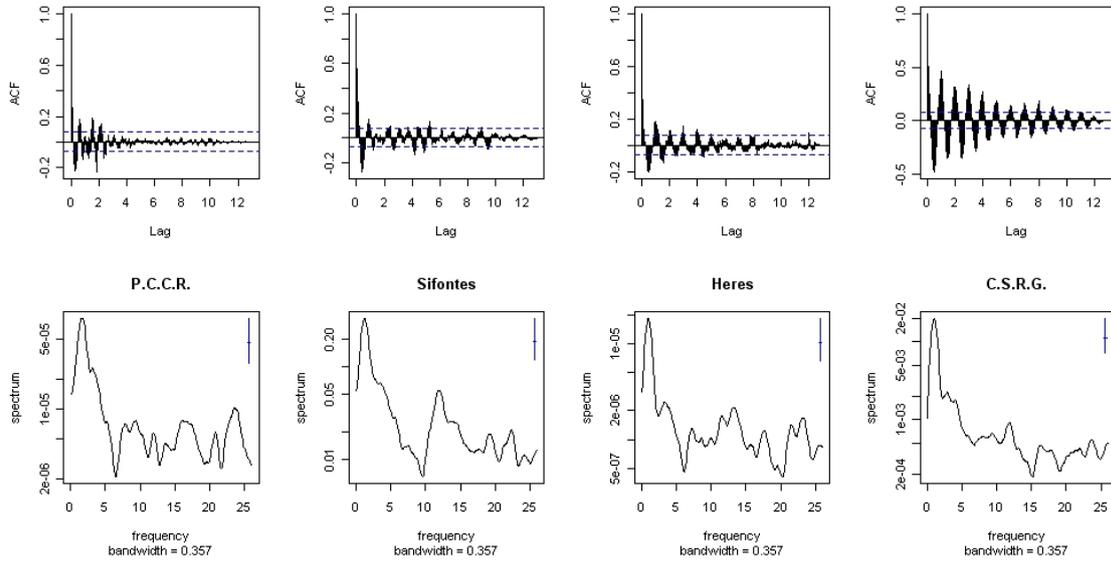


Fig. 7: Correlaciones Cruzadas de los sectores

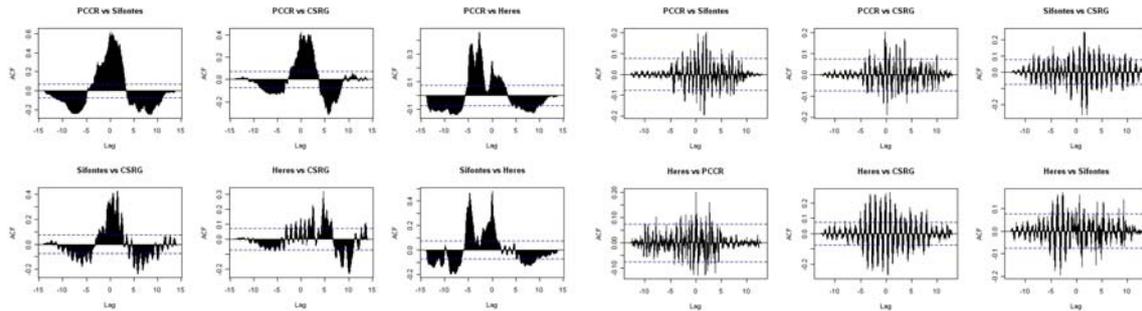
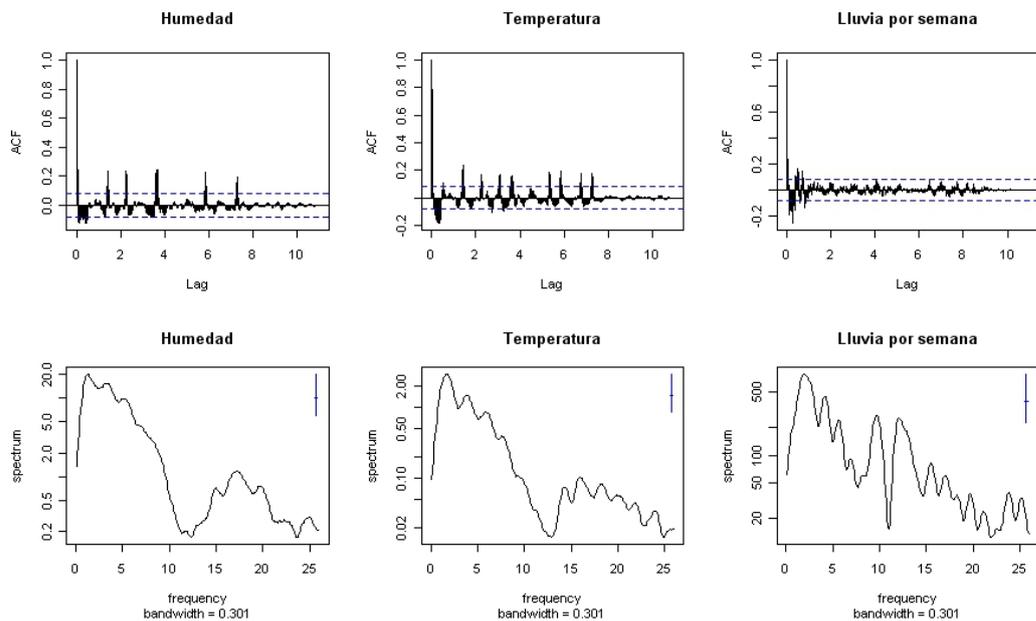


Fig. 8: ACF y densidad espectral de variables climáticas (Tumeremo, Bolívar)



Las variables climáticas registradas corresponden al área de Tumeremo, Municipio Sifontes. Se observó claramente que en la temperatura junto al ciclo anual existe un ciclo circa mensual semejante al que se ha encontrado repetidas veces en las series de malaria.

Por otro lado, en general la incidencia de malaria Sifontes y PCCR guarda gran similitud en su evolución con la lluvia y aunque en cuanto a su estructura no es tan claro. Pero es un indicio bastante claro de que la rigiere la dinámica de la malaria.

Fig. 9: Correlaciones Cruzadas entre IPA (Sifontes) y variables climáticas

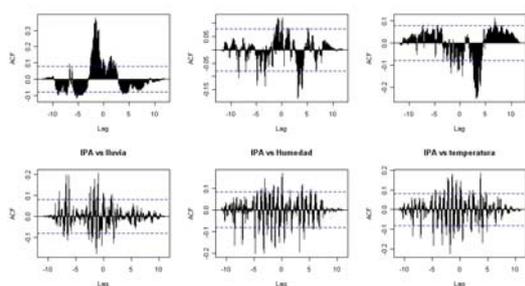
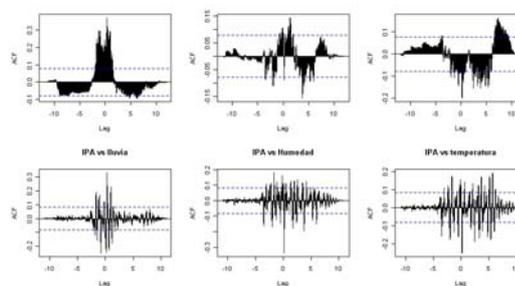
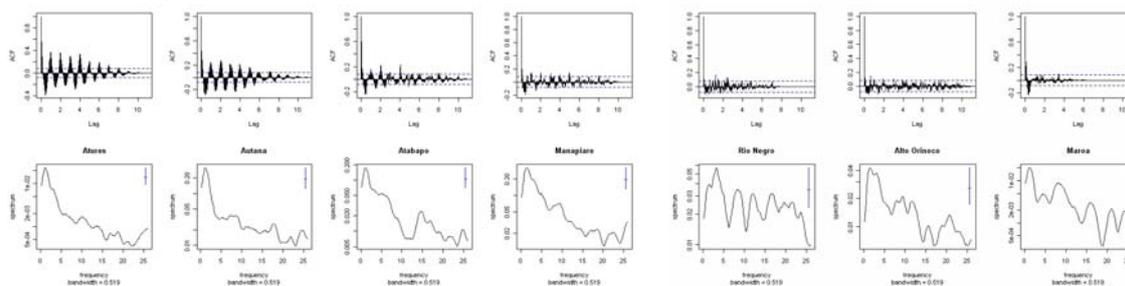


Fig. 10: Correlaciones Cruzadas entre IPA (Piar-Caróni) y variables climáticas



Amazonas

Fig. 12: ACF y densidad espectral por municipio en el estado Amazonas



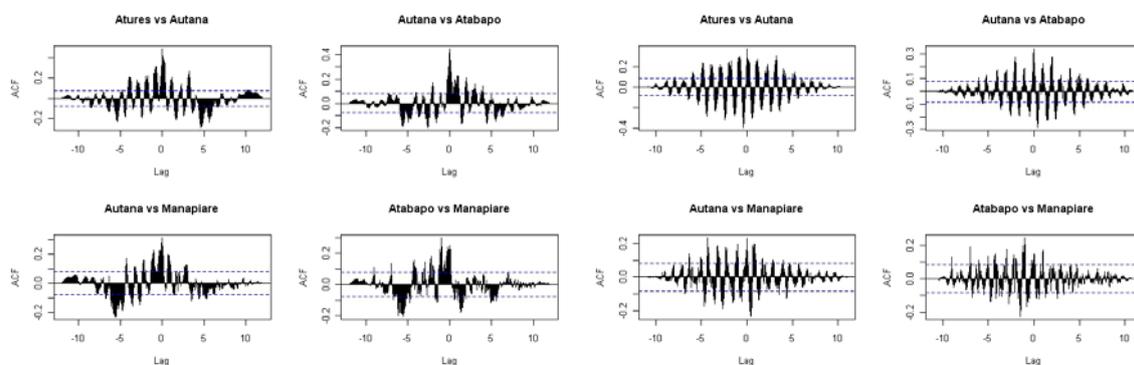
Los municipios Atures, Autana muestran una estructura estacional muy nítida, mientras que en Atabapo y Manapiare esta es algo mas irregular. En cambio Río Negro, Alto Orinoco y Maroa presentan una dinámica demasiado irregular para conseguir algo en limpio de ellas. Esto es muy posible que se deba a que son series de valores muy bajos y demasiados ceros (como se puede ver en la tabla 3), por lo que considero preferible de momento no trabajar con esas series (ver Fig. 14). Se observó que los

Tabla 3

	% ceros	IPA
Atures	0.0288	0,496
Manapiare	0.0593	1,913
Autana	0.0801	1,662
Alto Orinoco	0.1474	0,787
Atabapo	0.2099	1,028
Río Negro	0.7067	0,486
Maroa	0.9135	0,099

municipios Atures, Autana, Atabapo y Manapiare presentan una evolución hasta cierto punto semejante mientras que su estructura estacional es muy parecida (Fig 13).

Fig. 13: Correlaciones cruzadas entre los municipios Atures, Manapiare, Autana y Atabapo



IPA (con tendencia)

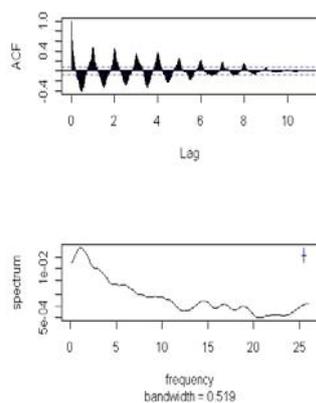
IPA (sin tendencia)

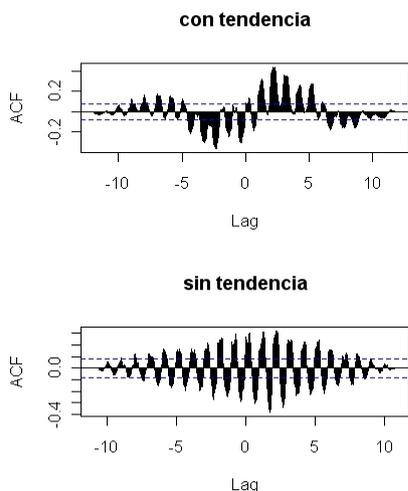
Sumando las series de Atures, Autana, Atabapo y Manapiare la serie resultante presenta una estructura estacional bastante clara (Fig 15). Al inspeccionarse la densidad se insinúa un pico a una frecuencia de 15/año (un periodo aprox. de 3.5 semanas)

Fig. 14: Municipios de amazonas. Solo se siguió trabajando con los municipios en gris oscuro.



Fig. 15: ACF y densidad espectral.





Comparando esta serie con la de CSRG se ve que si bien mantienen evoluciones distintas tiene una estructura estacional muy semejante

Conclusiones de la primera fase

- En general, en toda Guayana la incidencia de malaria presenta un comportamiento claramente estacional con un ciclo principal de duración anual y un ciclo secundario cuya duración varía en los distintos municipios entre 3 y 5 semanas.
- La estimación presenta problemas en algunos municipios del noreste de Bolívar y el sur de Amazonas debido a que poseen valores muy bajos y presentan muchos ceros. Aunque al sumar series semejantes se puede paliar un poco este problema.
- Se puede ver una muy clara influencia de la lluvia en la incidencia de malaria en el noreste de Bolívar. Es conveniente disponer de registros de lluvia de otras zonas para validar esta dependencia.

4.1.2. Segunda fase: identificación del Modelo ARIMAX.

En la segunda fase, mediante una selección paso a paso se escoge el modelo ARIMA óptimo para cada municipio, tomando como criterio de significancia el valor AIC. Este procedimiento se repitió incluyendo uno o mas regresores externos. Para los regresores externos se uso un solo rezago, de modo que el modelo ARIMAX toma la forma

$$\Phi(B)Y_t = \Theta(B)Z_t + \sum_{k=1}^b \xi_k W_{k,t-1}, \quad \{Z_t\} \sim RB(\sigma^2)$$

donde b es el numero de regresores externos y $Y_t = (1-B)^d X_t$. Los modelos óptimos obtenidos fueron a su vez comparados en base a su valor AIC y el p-valor de su test de Ljung-Box.

En Sucre, tanto para Benítez como para Cagigal se ajustaron modelos ARIMA muy semejantes, pero en ambos casos se presenta el problema de que el residuo no se puede considerar ruido blanco, lo cual le resta veracidad al modelo. Pero realizando una suma ponderada por su número poblacional de las series de Benítez y Cagigal se solucionó este problema

Al ajustar los modelos ARIMA agregando un regresor externo, se vio que el único caso en que mejoró la bondad de ajuste es con la inclusión de la variable

Tabla 4: Comparación de los modelos ARIMA. Estado Sucre

	Modelo ARIMA	Regresor(es) externo(s)	AIC	LB.test p-valor
Benítez	ARIMA(5, 1, 2)(2, 0, 1)[52]		-639,39	0,04
Cagigal	ARIMA(5, 1, 3)(2, 0, 1)[52]		-408,75	0,01
Paria (1)	ARIMA(3, 1, 4)(2, 0, 1)[52]		-712,12	0,14
Paria (2.1)	ARIMA(5, 1, 5)(2, 0, 1)[52]	Precipitación	-756,07	0,11
Paria (2.2)	ARIMA(2, 1, 1)(2, 0, 1)[52]	Temperatura	-677,78	0,32
Paria (2.3)	ARIMA(2, 1, 1)(2, 0, 2)[52]	Humedad	-673,77	0,21
Paria (2.4)	ARIMA(2, 1, 1)(2, 0, 1)[52]	Pr, Tm, Hm	112,96	0,00

precipitación. De hecho cuando se ensaya incluir tanto precipitación, temperatura como humedad el residuo del modelo ajustado no pasa un test de Ljung-Box y la bondad de ajuste cae notablemente

En Amazonas al comparar los ARIMA ajustados sin ningún regresor con los ajustados tomando la precipitación (PP) como regresor externo, se notó que agregar la precipitación como regresor externo mejoró la bondad de ajuste del modelo, salvo el municipio Atures. Estos resultados confirmaron lo observado en el estado Sucre

Tabla 5: Comparación de los modelos ARIMA. Estado Amazonas

	Modelo ARIMA	Xreg	AIC	LB.test p.valor
Atabapo (1)	ARIMA(3, 1, 4)(0, 0, 1)[52]		1727,73	0,11
Atabapo (2)	ARIMA(3, 1, 3)(0, 0, 2)[52]	PP	1614,20	0,08
Atures (1)	ARIMA(2, 1, 4)(0, 0, 1)[52]		29,39	0,26
Atures (2)	ARIMA(2, 1, 4)(0, 0, 1)[52]	PP	63,13	0,24
Autana (1)	ARIMA(1, 1, 1)(0, 0, 2)[52]		1945,00	0,11
Autana (2)	ARIMA(1, 1, 1)(0, 0, 2)[52]	PP	1818,65	0,22
Manapiare (1)	ARIMA(2, 1, 2)(0, 0, 0)[52]		2191,31	0,80
Manapiare (2)	ARIMA(2, 1, 2)(0, 0, 0)[52]	PP	2050,52	0,86
Maroa (1)	ARIMA(3, 1, 3)(0, 0, 0)[52]		496,66	0,00
Maroa (2)	ARIMA(3, 1, 3)(0, 0, 0)[52]	PP	485,24	0,00
Alto Orinoco (1)	ARIMA(2, 1, 2)(0, 0, 2)[52]		1543,06	0,13
Alto Orinoco (2)	ARIMA(2, 1, 2)(0, 0, 0)[52]	PP	1435,17	0,17
Río Negro (1)	ARIMA(2, 1, 3)(0, 0, 0)[52]		1998,27	0,37
Río Negro (2)	ARIMA(2, 1, 3)(0, 0, 0)[52]	PP	1737,87	0,08

En el caso de Bolívar los resultados no son tan homogéneos. Ya que por un lado, en los municipios Sifontes, Gran Sabana, Raúl Leoni y Roscio al agregar la precipitación como regresor externo aumenta la bondad de ajuste para el ARIMA hallado, lo cual no ocurre en los municipios Piar, Cedeño y Sucre.

Fig. 14: Municipios de Bolívar agrupados en base a la similitud de sus dinámicas.



Tabla 6: Comparación de los modelos ARIMA. Estado Sifontes

Municipio	Modelo ARIMA	Xreg	χ^2	AIC	LB.test p.valor
Sifontes (1)	ARIMA(5, 1, 5)(0, 0, 0)[52]		1,09	1999,72	0,15
Sifontes (2)	ARIMA(5, 1, 4)(0, 0, 0)[52]	PP	1,19	1898,95	0,23
Roscio (1)	ARIMA(5, 1, 4)(0, 0, 0)[52]		0,58	1575,01	0,00
Roscio (2)	ARIMA(5, 1, 4)(0, 0, 0)[52]	PP	0,63	1507,07	0,01
Piar (1)	ARIMA(0, 1, 1)(0, 0, 2)[52]		0,03	-377,57	0,00
Piar (2)	ARIMA(5, 1, 4)(0, 0, 0)[52]	PP	0,03	-307,64	0,00
Gran Sabana (1)	ARIMA(5, 0, 5)(2, 0, 1)[52]		0,25	1005,50	0,03
Gran Sabana (2)	ARIMA(4, 1, 3)(2, 0, 2)[52]	PP	0,20	782,44	0,08
Raúl Leoni (1)	ARIMA(2, 1, 1)(1, 0, 0)[52]		0,24	971,34	0,18
Raúl Leoni (2)	ARIMA(2, 1, 1)(1, 0, 0)[52]	PP	0,22	848,77	0,22
Sucre (1)	ARIMA(1, 1, 2)(0, 0, 2)[52]		0,12	495,55	0,05
Sucre (2)	ARIMA(1, 1, 2)(0, 0, 2)[52]	PP	0,13	509,64	0,10
Cedeño (1)	ARIMA(3, 1, 4)(0, 0, 1)[52]		0,05	-119,07	0,00
Cedeño (2)	ARIMA(4, 1, 1)(1, 0, 1)[52]	PP	0,05	-71,98	0,02

Es de hacer notar que en aquellos municipio donde no mejoro la bondad de ajuste al incluir la precipitación como regresor externo, tampoco el residuo del modelo ARIMA paso el test de Ljung-Box (con la excepción de Sucre donde si bien el residuo se distribuye como ruido blanco, no mejoró la bondad de ajuste,

Tabla 7: Diferencia en el p-valor del test de Ljung-Box dado que se incluya o no la precipitación como variable regresora

municipio	regresor externo		Δ p.valor
	no	PP	
Sifontes	0,15	0,23	0,07
Roscio	0,00	0,01	8,E-03
Piar	0,00	0,00	7,E-07
Gran Sabana	0,03	0,08	0,05
Raúl Leoni	0,18	0,22	0,04
Sucre	0,05	0,10	0,05
Cedeño	0,00	0,02	0,02

mientras que Roscio a pesar de no pasar el test de Ljung-Box, se redujo el valor AIC). Si se miran con mayor detenimiento los p-valores del test de Ljung-Box, se hace notorio que los p-valores del test aumentan al agregar la precipitación como regresor externo (tabla 7) o sea que la distribución del residuo se aproxima mas a la del ruido blanco. Esto puede interpretarse claramente como que la precipitación mejora el ajuste.

4.1.3. Tercera fase: estimación de los parámetros del modelo ARMAX

En la tercera fase, se ajustaron los datos a un modelo ARMAX, un caso particular de los ARIMAX en el que el orden d es igual a 0. Por otro lado, el orden r usado no es necesariamente igual a 1. por lo que el modelo toma la forma

$$\Phi(B)X_t = \Theta(B)Z_t + \sum_{k=1}^b \Xi_k(B)W_{k,t}, \quad \{Z_t\} \sim RB(\sigma^2)$$

donde $\Xi(z) = 1 - \xi_1 z - \dots - \xi_r z^r$. La metodología usada en esta fase se puede resumir en los siguientes pasos:

1. Determinación de r óptimo; esto se hace mediante el procedimiento para calcular coeficientes β_k con distintos R^2 , considerándose el r óptimo aquel que arroje el mayor R^2 . el rango que fue usado iba de 0 a 104 rezagos. (este paso no es estrictamente necesario)
2. Estimación de los coeficientes β_k mediante una regresión lineal múltiple.
3. Calculo de pY_t a partir de los coeficientes β_k .
4. Crear serie ajustada W_t ($W_t = Y_t - pY_t$)
5. Estimar el mejor modelo ARMA según el BIC para la serie original (Y_t) así como para la serie ajustada (W_t)
6. Compara los dos modelos ARMA ajustados en base a su BIC y el p-valor del test de Ljung-Box.

En base a los resultados obtenidos en la fase previa se opto por usar solamente las siguientes series:

- Paria (suma ponderada de Benítez y Cagigal) del año 2003 al 2009, tomando como variables regresoras tentativas a la precipitación, la temperatura y la humedad.
- Atures (edo. Amazonas) del año 1999 al 2009, con la precipitación y el nivel de río como variables regresoras tentativas.
- Sifontes (edo. Bolivar) del año 1999 al 2009 y posteriormente del 2004 al 2009. las variables regresoras tentativas usadas fueron la precipitación, la temperatura y la humedad.

Determinación de r óptimo

Para la serie Paria (2003-2009), al analizar las curvas R^2 vs r , se observó que dentro del rango escogido no se alcanza ningún máximo para R^2 , de modo que se podría suponer que la influencia del clima es de muy largo plazo. Por otro lado se puede ver que para cualquier r , la temperatura y la humedad actúan como mejores regresores que la precipitación.

En tanto que con la serie Atures (1999-2009) se observó lo que ya había encontrado en Paria: dentro del rango escogido de r no se alcanza ningún máximo para R^2 . Por otro lado se puede ver que el nivel de río no es tan buen regresor externo como la precipitación.

Con la serie Sifontes (1999-2009), las curvas de R^2 vs r concuerdan con los resultados anteriores respecto a que R^2 no alcanza ningún máximo dentro del rango escogido de r . En lo que si difiere de los resultados anteriores es que la curva correspondiente a la precipitación se corta con correspondientes a temperatura y humedad.

Pero en cambio con Sifontes (2004:2009) los resultados son muy semejantes a los que ya se habían visto en Paria, por lo que la irregularidad antes reportada se debe atribuir más que nada al cambio de nivel.

Fig 17: coeficiente de determinación R^2 vs rezago máximo r .

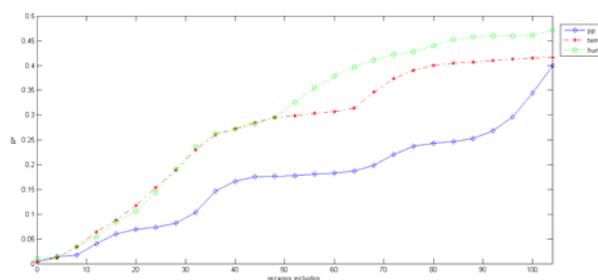


Fig 18: coeficiente de determinación R^2 vs rezago máximo r .

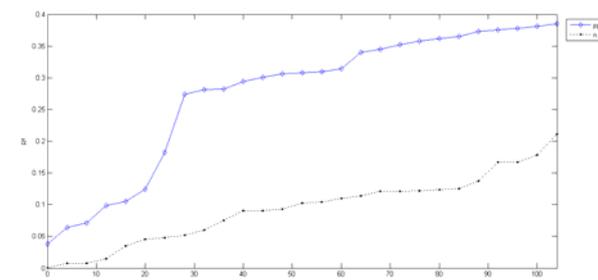


Fig 19: coeficiente de determinación R^2 vs rezago máximo r .

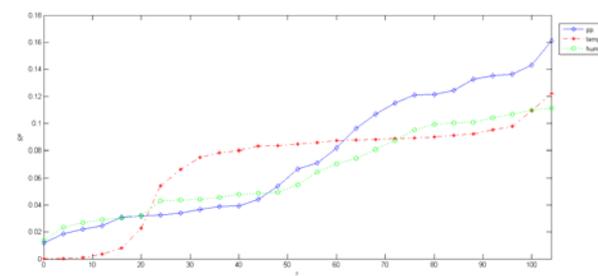
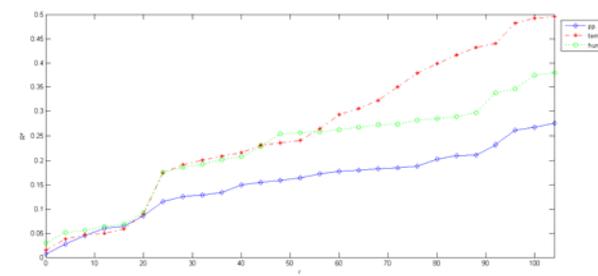


Fig. 20: coeficiente de determinación R^2 vs rezago máximo r .



Ajuste a un Modelos ARMAX

En la siguiente tabla se presenta un resumen de los resultados de los ajustes de las tres series ya mencionadas a modelos ARMAX. Contiene los siguientes campos:

- W : fórmula de la serie ajustado. Con t esta referido la temperatura, con h la humedad, con p la precipitación y con r el nivel de río
- R^2 : coeficiente de determinación de la regresión múltiple (estimación de coeficiente β_k)
- σ^2 : (varianza del residuo del ARMA ajustado).
- LB: p-valor del test de Ljung-Box.
- p, P, q, Q : ordenes del ARMA ajustado.
- k : número de parámetros del modelo ($p + P + q + Q + 1$)

Tabla 8: resumen de resultados.

		R^2	σ^2	LB	BIC	p	P	q	Q	k
Paria (2003-2009)	Y		0,006	9,30E-05	-678,6	2	2	2	0	7
	$Y - p(Y t)$	0,42	0,006	7,20E-04	-682,51	2	2	0	0	5
	$Y - p(Y h)$	0,47	0,012	1,90E-01	-457,86	1	2	2	0	6
	$Y - p(Y p)$	0,4	0,009	1,40E-01	-557,92	2	2	1	0	6
Atures (1999-2009)	Y		0,064	2,70E-01	81,19	1	1	1	0	4
	$Y - p(Y p)$	0,39	0,054	2,90E-01	-13,92	1	1	2	0	5
	$Y - p(Y r)$	0,21	0,058	7,50E-02	18,74	2	1	0	0	4
Sifontes (1999-2009)	Y		1,506	3,30E-02	1716,59	1	0	2	0	4
	$Y - p(Y t)$	0,12	1,193	3,00E-07	1594,56	1	0	2	0	4
	$Y - p(Y h)$	0,11	1,101	4,00E-08	1552,94	1	0	2	0	4
	$Y - p(Y p)$	0,16	1,433	9,70E-12	1690,6	1	0	1	1	4
Sifontes (2004-2009)	Y		1,271	3,30E-02	851,03	5	2	0	0	8
	$Y - p(Y t)$	0,5	1,738	5,60E-07	904,44	1	2	0	0	4
	$Y - p(Y h)$	0,38	1,626	7,60E-04	899,36	1	2	0	0	4
	$Y - p(Y p)$	0,28	2,807	5,10E-09	1033,19	1	2	0	0	4

Se observó que en Paria la aplicación de la modelización ARMAX es satisfactoria en la medida que el p-valor del test de Ljung-Box se redujo (aunque esa reducción es bastante pequeña con la temperatura como regresor externo), lo cual hace que el ARMA estimado sea mucho más fiable en sus pronósticos. De hecho el residuo del ARMA ajustado no se puede considerar ruido blanco. En cambio el BIC favoreció la temperatura como regresor externo debido sobre todo a que con la temperatura el orden total se redujo y no aumentó σ^2 . Con lo que es conveniente dejar en claro que cuando

hay contradicciones entre lo el criterio BIC y el criterio Ljung-Box tiene precedencia el criterio Ljung-Box.

En Atures, se observó en principio que el residuo del ARMA ajustado a la serie original se puede considerar ruido blanco y al usar la precipitación como regresor externo el p-valor del test de Ljung-Box apenas cambio, mientras que al usar el nivel de río como regresor externo mas bien redujo el p-valor del test de Ljung-Box, lo que indica que el nivel de río no es una buena variable explicativa. Lo que si es notorio es que en ambos casos el BIC se redujo en ambos casos. Esto a causa de que con ambos regresores se redujo σ^2 .

En Sifontes (1999-2008), en cambio el uso de de regresores externos redujo sensiblemente el p-valor del test de Ljung-Box, lo cual mas bien empeoró la validez de cualquier pronostico, ya que el p-valor en la serie original sólo es moderadamente bajo. Aunque el BIC se redujo con cualquiera de las variables usadas como regresor externo ya que para todos los casos σ^2 se redujo. Se había supuesto que tal vez el cambio de nivel tan evidente que se observaba en la serie podría haber afectado las estimaciones, de modo que se usó una serie mas corta abarcando desde 2004 (año donde se sitúa el cambio de nivel) hasta 2008. El hecho es que no solo se reduce el p-valor del test de Ljung-Box sino que además aumenta el BIC, debido a que es incrementado por el incremento de σ^2 a pesar de que hubo una reducción drástica del numero de parámetros del modelo respecto al ARMA ajustado a la serie original. Es muy posible que las factores antrópicas (en particular, la actividad minera en el área) sean tan o mas preponderantes que las variables climáticas en la dinámica de malaria.

Conclusiones de la tercera fase

- Todo indica que la influencia de las variables climáticas en la dinámica de la malaria (IPS) es de largo plazo, de modo que es recomendable que se use un r considerablemente grande.
- La temperatura y la humedad son las variables climáticas que mejor responden como variables regresoras, mientras el nivel de río es una mala variable regresora.
- Es necesario tomar en cuenta la relevancia de factores antrópicas ya que si estos son demasiado preponderantes para que esta metodología no arroje resultados

satisfactorios, ya que toma como supuesto que la tendencia de la dinámica de malaria puede ser explicado por alguna variable climática.

- El orden del ARMA, bien sea p o q , no excede 2 salvo en un caso. Esto posiblemente refleja el lapso durante el cual se presentan los primeros síntomas, como es referido en la descripción del ciclo de vida del Plasmodium.

4.2 Resultados.

4.2.1. Selección de modelos

Recapitulando, la influencia de las variables exógenas se expresa mediante un modelo de retardos distribuidos $RD(r_1, \dots, r_b)$, donde r_i el número máximo de retardos para la variable regresora $\{x_{i,t}\}$. Como variables regresoras tentativas a ser usadas en el modelo de retardos distribuidos se dispone de las siguientes: Precipitación acumulada (P_h), Humedad (H) y Temperatura media (T).

Los resultados del muestreo del espacio paramétrico, los cuales están resumidos en las tablas 9 y 10. Cada tabla contiene los siguientes estadísticos:

- r : Máximo número de rezagos de la variable regresora. (en el caso del modelo bivariado (r_p corresponde al rezago máximo asignado a la precipitación acumulada y r_k al rezago máximo asignado a la otra variable regresora))
- h : Máximo número de rezagos usada para la precipitación acumulada.
- R^2 : Coeficiente de determinación.

Tabla 9: Sumario – modelos univariados

Benítez				Cagigal			
modelo de regresión	r	h	R^2	modelo de regresión	r	h	R^2
IPS $\sim P_h$	80	26	0,133	IPS $\sim P_h$	56	22	-0,095
IPS $\sim H$	77	—	-0,108	IPS $\sim H$	78	—	-0,135
IPS $\sim T$	78	—	-0,102	IPS $\sim T$	78	—	-0,201
$\log(\text{IPS}+1) \sim \log(P_h+1)$	80	17	-0,200	$\log(\text{IPS}+1) \sim \log(P_h+1)$	80	21	-0,260
$\log(\text{IPS}+1) \sim \log(H)$	77	—	-0,113	$\log(\text{IPS}+1) \sim \log(H)$	78	—	-0,192
$\log(\text{IPS}+1) \sim \log(T)$	78	—	-0,088	$\log(\text{IPS}+1) \sim \log(T)$	78	—	-0,236

Tabla 10: Sumario – modelos bivariados

Benítez					Cagigal				
Modelo	r_k	r_p	h	R^2	Modelo	r_k	r_p	h	R^2
IPS $\sim P_h + H$	64	80	26	0,394	IPS $\sim P_h + H$	64	76	26	0,332
IPS $\sim P_h + T$	80	28	26	0,358	IPS $\sim P_h + T$	80	64	17	0,320
$\log(\text{IPS} + 1) \sim \log(P_h + 1) + \log(H)$	72	44	8	0,558	$\log(\text{IPS} + 1) \sim \log(P_h + 1) + \log(H)$	80	72	13	0,463
$\log(\text{IPS} + 1) \sim \log(P_h + 1) + \log(T)$	76	80	13	0,503	$\log(\text{IPS} + 1) \sim \log(P_h + 1) + \log(T)$	80	64	14	0,403

De estos resultados se pueden extraer las siguientes conclusiones:

- Los ajustes óptimos se obtuvieron para valores de r muy altos, de hecho, cercanos al borde como se ve en las figura 21 y 22. Esto implicaría que el máximo se podría encontrar fuera del intervalo explorado.
- Con los modelos bivariados se obtuvo un mejor ajuste que con los univariados. El mayor R^2 obtenido con los modelos univariados fue de 0.26 por debajo del obtenido con cualquiera de los bivariados, todos por encima de 0.32.
- al aplicarse la transformada logarítmica se obtiene un mejor ajuste. Mientras que con los modelos bivariados no logarítmicos el R^2 esta entre 0.3 y 0.4, con los modelos bivariados logarítmicos este se sitúa por encima de 0.4.
- Además que con la transformada logarítmica el h óptimo obtenido es menor que el que se obtuvo para los no logarítmicos.

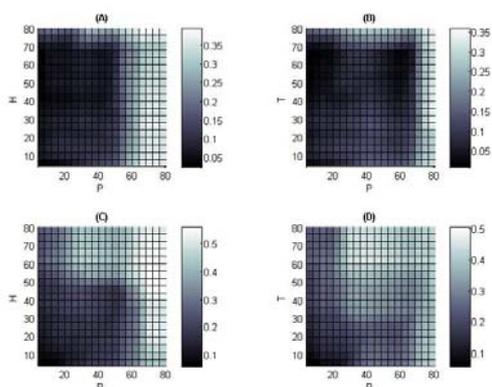


Figura 21: mapa paramétrico – municipio Benítez

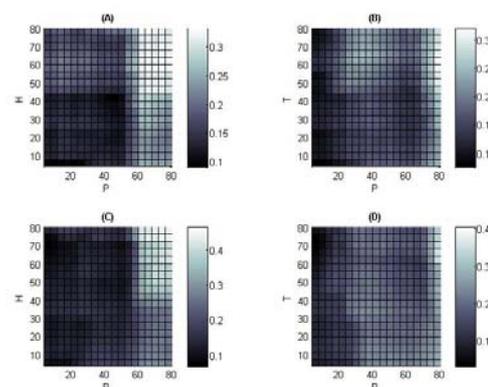


Figura 22: mapa paramétrico – municipio Cagigal

En los mapas paramétricos se muestran los coeficientes de determinación (R^2) para distintos numero máximo de retardos (r_t) correspondiente a Precipitación (P), Temperatura (T) y Humedad (H) aplicado a los siguientes modelos:

- a) $IPS \sim P_h + H$ c) $\log(IPS + 1) \sim \log(P_h + 1) + \log(H)$
 b) $IPS \sim P_h + T$ d) $\log(IPS + 1) \sim \log(P_h + 1) + \log(T)$

Como ya se dijo, el orden óptimo podría estar fuera del intervalo explorado. Pero usar un r muy alto aparte de ser computacionalmente costoso, limita la longitud de la serie de la variable respuesta (IPS) que puede ser utilizada, ya que el inicio de ésta debe ser r semanas posterior al inicio de la variable regresora. De modo que por razones prácticas, en adelante se usara $r = 52$, para cualquiera de las variables regresoras.

Aunque los resultados anteriormente mostrados, permiten suponer que los modelos logarítmicos bivariados resultan mejores, para dar mayor peso a esta elección, se ajustó un modelo autorregresivo al residuo de los modelos ya mencionados, usando el registro del municipio Benítez. Estos autorregresivos se compararan según su valor BIC y el p-

valor de su test de Ljung-Box, de modo de asegurar que el residuo del autorregresivo se ajuste a ruido blanco. De los resultados obtenidos (los cuales están resumidos en la tabla 11) se pueden extraer las siguientes conclusiones:

- Es de hacer notar que para cualquiera de los modelos el modelo autorregresivo obtenido mediante selección paso a paso fue invariablemente de orden 2.
- Con los modelos logarítmicos se obtiene un mejor ajuste (como indica el criterio BIC), ya que el residuo del autorregresivo posee una varianza mas reducida.
- Con los modelos bivariados el residuo del autorregresivo siempre se ajustaba razonablemente a ruido blanco, mientras que con los univariados en unos casos se obtenía un buen ajuste en otros casos no, independiente de si era logarítmico o no.

Todo esto confirma la conclusión anterior de que los modelos logarítmicos bivariados son los más idóneos para representar la influencia de las variables exógenas. Por lo que sólo bastaría con comparar los coeficientes de determinación de los modelos de retardos distribuidos para seleccionar el modelo. Además queda claro que el mejor modelo para representar al residuo es un autorregresivo de orden 2.

Tabla 11: Ajuste de los residuos del modelo de retardos distribuidos a un modelo autorregresivos.

MODELO	R ²	p	σ^2	BIC	p-valor del test de L.B.
IPS ~ P	0,173	2	1,4E-02	-432,04	0,007
IPS ~ H	0,307	2	1,3E-02	-443,97	0,021
IPS ~ T	0,254	2	1,4E-02	-419,12	0,205
LOG(IPS + 1) ~ LOG(P)	0,241	2	6,2E-03	-681,70	0,093
LOG(IPS + 1) ~ LOG(H)	0,344	2	6,4E-03	-670,17	0,006
LOG(IPS + 1) ~ LOG(T)	0,352	2	6,9E-03	-649,39	0,118
IPS ~ P + H	0,454	2	1,2E-02	-471,33	0,057
IPS ~ P + T	0,361	2	1,2E-02	-471,33	0,057
LOG(IPS + 1) ~ LOG(P + 1) + LOG(H)	0,491	2	6,0E-03	-694,97	0,148
LOG(IPS + 1) ~ LOG(P + 1) + LOG(T)	0,472	2	5,5E-03	-720,88	0,115

Ya comprobado que el modelo logarítmico bivariado (precipitación + temperatura) es con el que se logra el mejor ajuste, queda por discernir cual es la medida de temperatura mas idónea. Se disponen de 4 medidas tentativas: Temperatura mínima semanal, máxima semanal y media semanal en base al registro de la estación meteorológica de la aviación en Güiría, municipio Valdez. Además de la temperatura media semanal de la superficie del mar (El Niño). Los modelos con las distintas medidas de temperatura se compararon mediante su coeficiente de determinación (R²). Se encontró que la medida de temperatura con la que se obtiene el mejor ajuste es con El Niño, como se puede ver

en la tabla 4. Mientras, la temperatura máxima resulto el mejor regresor después de El Niño. De hecho con ninguna de los registros de Güiria se obtiene un R^2 mayor a 0.5.

Como se puede ver en los municipios mas alejados del Guiria (municipio. Valdez) se obtiene un ajuste bastante pobre. Esto se puede deber muy posiblemente a la gran cantidad de ceros que hay en estos registros. Además, estos municipios son los más alejados de la estación de Güiria, donde se registra tanto la temperatura como la precipitación (la cual es usada en cualquiera de los casos). De modo que para posteriores análisis,

Tabla 12: coeficientes de determinación del modelo de retardo distribuido según la medida de temperatura usada.

Localidad	El Niño	T. Med.	T. Min	T. Max	Mejor Regresor
A. E. Blanco	0,458	0,266	0,270	0,384	niño
A. Mata	0,539	0,152	0,183	0,231	niño
Arismendi	0,430	0,215	0,224	0,303	niño
Benítez	0,522	0,255	0,244	0,341	niño
Bermúdez	0,321	0,337	0,314	0,352	max
Bolivar	0,290	0,357	0,336	0,332	med
C. S. Acosta	0,213	0,322	0,260	0,278	med
Cajigal	0,538	0,259	0,215	0,398	niño
Libertador	0,372	0,286	0,275	0,366	niño
Mariño	0,517	0,322	0,235	0,451	niño
Mejía	0,235	0,331	0,297	0,289	med
Montes	—	—	—	—	
Ribero	0,398	0,367	0,288	0,470	max
Sucre	0,410	0,302	0,287	0,316	niño
Valdez	0,421	0,355	0,280	0,438	max
rango	0,213	0,152	0,183	0,231	
	0,539	0,367	0,336	0,470	

quedarían excluidos los municipios C. S. Acosta, Mejía, Montes, Bolívar y Sucre. Adicionalmente se excluyen por no tener registro alguno, a las parroquias: San José de Aerocuar (Andrés Mata), General Francisco Antonio Vásquez y Tunapuicito (Benítez), Macarapana, Santa Rosa y Santa Teresa (Bermúdez), Santa Cruz y Villa Frontado (Ribero) y Santa Fe (Sucre). Con lo que seria excluidas, 17 parroquias, con lo que el set de datos queda reducido a 31 parroquias.

4.2.2. Optimización de parámetros del modelo mediante MCMC

Una vez seleccionado el modelo, se estimaron los parámetros del modelo mediante un algoritmo MCMC.

Tabla 13: índices del modelo antes y después de la optimización

Mun	Par	Varianza del residuo z		p-valor del test LB		p-valor del test F	
		inicial	opt	inicial	opt	inicial	opt
A. E. Blanco	Mariño	9,1E-08	9,1E-08	0,00	0,00	0,00	0,00
A. E. Blanco	R. Gallegos	9,1E-08	9,1E-08	0,00	0,00	0,00	0,00
A. Mata	T. Acosta	9,1E-08	9,1E-08	0,00	0,00	0,00	0,00
Arismendi	A. J. de Sucre	9,1E-08	1,3E-06	0,00	0,12	0,00	0,00
Arismendi	M. de P. Santo	1,3E-06	4,5E-09	0,00	0,00	0,00	0,06
Arismendi	P. Santo	4,5E-09	1,7E-08	0,00	0,00	0,00	0,00
Arismendi	Río Caribe	1,7E-08	1,7E-08	0,00	0,00	0,00	0,00
Arismendi	S. J. Galdonas	1,7E-08	1,5E-06	0,00	0,00	0,00	0,00
Benítez	El Pilar	1,5E-06	1,5E-06	0,00	0,00	0,00	0,00
Benítez	El Rincón	2,5E-10	3,2E-09	0,00	0,00	0,00	0,00
Benítez	Guaraúnos	3,2E-09	1,1E-07	0,00	0,00	0,00	0,00
Benítez	Unión	1,1E-07	7,4E-07	0,00	0,00	0,00	0,00
Bermúdez	Bolívar	7,4E-07	7,4E-07	0,00	0,00	0,00	0,00
Bermúdez	Santa Catalina	7,4E-07	4,0E-09	0,00	0,00	0,00	0,00
Cajigal	El Paujil	4,0E-09	4,3E-07	0,00	0,02	0,00	0,00
Cajigal	Libertad	4,3E-07	4,8E-07	0,00	0,00	0,00	0,00
Cajigal	Yaguaraparo	4,8E-07	1,8E-07	0,00	0,00	0,00	0,01
Libertador	Campo Elías	3,8E-07	8,5E-07	0,02	0,25	0,00	0,09
Libertador	Tunapuy	8,5E-07	1,4E-08	0,00	0,00	0,00	0,00
Mariño	Campo Claro	1,4E-08	7,0E-09	0,00	0,00	0,00	0,00
Mariño	Irapa	7,0E-09	7,0E-09	0,00	0,00	0,00	0,00
Mariño	Marabal	7,0E-09	2,3E-07	0,00	0,00	0,00	0,24
Mariño	S. A. Irapa	2,3E-07	3,1E-08	0,00	0,00	0,00	0,00
Mariño	Soro	3,1E-08	1,2E-06	0,00	0,05	0,00	0,36
Ribero	Catuaro	1,2E-06	1,2E-06	0,00	0,00	0,00	0,00
Ribero	Rendón	1,2E-06	3,3E-07	0,00	0,07	0,00	0,00
Ribero	Santa María	3,3E-07	4,4E-09	0,00	0,00	0,00	0,00
Valdez	Bideau	4,4E-09	1,3E-06	0,00	0,00	0,00	0,00
Valdez	C. Colón	1,3E-06	1,1E-05	0,00	0,17	0,00	0,08
Valdez	Güiria	1,1E-05	1,1E-05	0,00	0,00	0,00	0,00
Valdez	P. Piedra	1,1E-05	2,9E-07	0,00	0,11	0,00	0,61

En la tabla 13 se muestran los índices del modelo antes y después de la optimización, es decir Varianza del residuo z, p-valor del test de Ljung-Box y p-valor del test F. aquellos p-valor que están por encima de 0.05 (alpha convencional de rechazo) están resaltados en negrita.

La optimización genera un incremento en los p-valores de los test de Ljung-Box y del test F, haciendo menos probable que sean falsas las hipótesis nulas (es decir que el ruido sea no correlacionado entre si y con las variables explicativas. Pero solo en un tercio de los municipios este incremento es apreciable en al menos uno de los indicadores y solo cuatro municipios donde hubiese un incremento notorio en ambos indicadores, como se ve mas claramente en la figura 23.

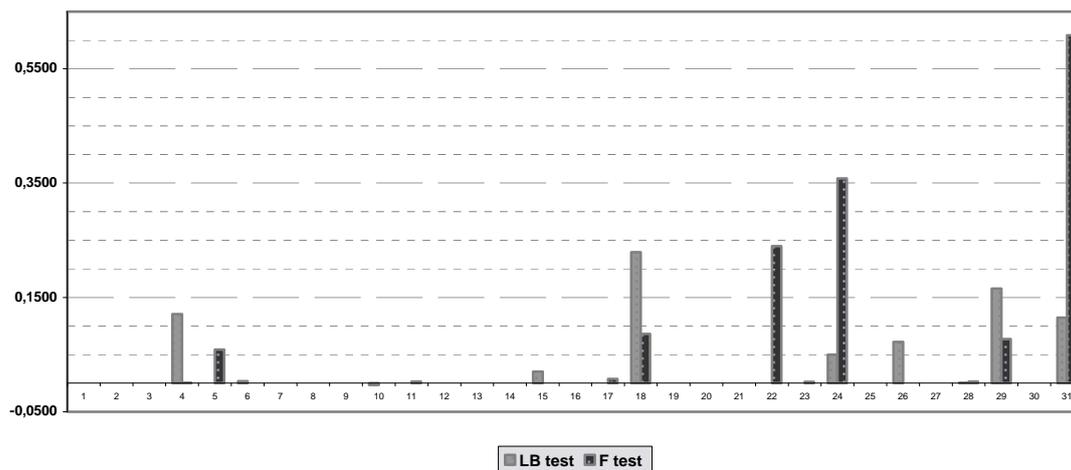


Fig. 23: Δp -valor de los test de Ljung-Box y F

Si bien el método de optimización tiende a causar un incremento en la varianza del residuo, hay que recalcar

que existe una correlación negativa entre el incremento de la varianza y el incremento de los p-valores del test F y del test de Ljung-Box. Esto se puede ver claramente en la tabla 14, la cual muestra la matriz de covarianza del cociente de la varianza del residuo del modelo optimizado y la varianza del modelo no optimizado (qV) y las diferencias de los p-valores del test de Ljung-Box (ΔLB) y el test F (ΔF) antes y después de la optimización. Todos estos datos fueron estandarizados antes de calcularse su matriz de covarianza.

Los coeficientes temporales ϕ presentan una tendencia descendente, haciéndose cada vez mas negativos conforme aumenta su índice. Todos los coeficientes espaciales λ (excepto el coeficiente 14 correspondiente a la parroquia Santa Catalina, municipio Bermúdez) toman valores entre -10 y 10.

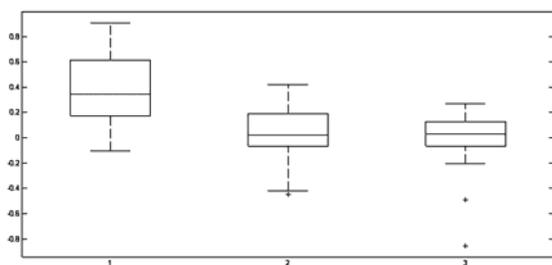


Fig. 24: diagramas de caja de los coeficientes temporales ϕ .

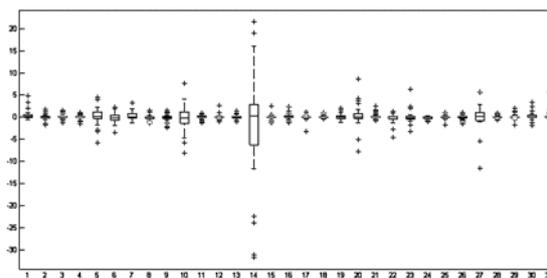


Fig. 25: diagramas de caja de los coeficientes espaciales λ .

Los coeficientes espaciales tienen una distribución cuya forma se mantiene independiente del índice que se escoja. Para cualquier índice se encuentra una media cercana a una desviación standard y el rango se ubica entre cuatro y seis desviaciones standards como se ve en la figura 27.

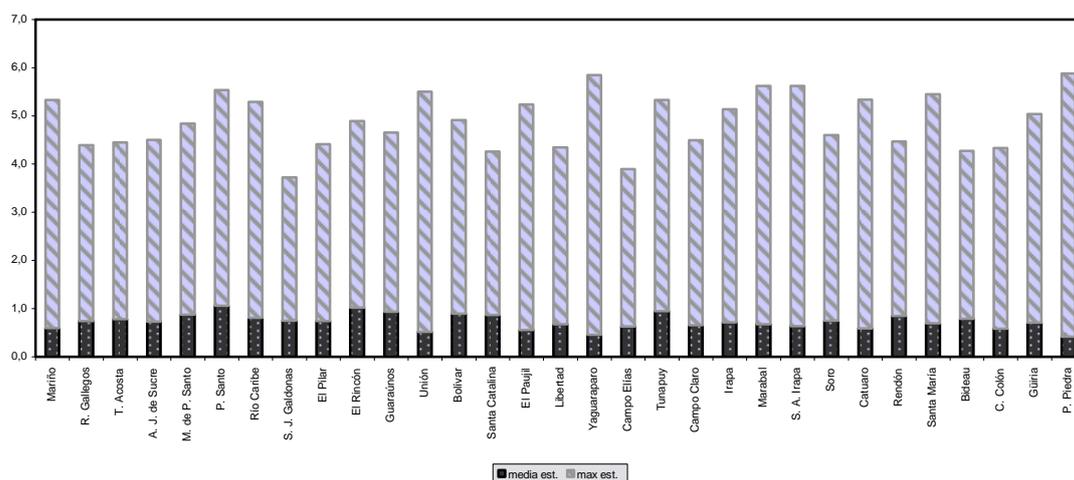


Fig. 26. media y valor máximo de los valores absolutos de los coeficientes espaciales l estandarizados por la desviación standard

La media de los valores absolutos de los coeficientes espaciales λ se puede ser usada para discernir con mayor claridad la influencia de una localidad sobre el resto de la vecindad. En la figura 28, se puede ver como la parroquia Santa Catalina (índice 14) tiene un influencia muy marcada sobre la dinámica de las demás parroquias. Además las parroquias M. de P. Santo (5), P. Santo (6) y Río Caribe (7) del municipio Arismendi, El Rincón (10), municipio Benítez, Campo Claro (20), municipio Mariño y Santa María (27), municipio Ribero tienen una influencia moderadamente alta.

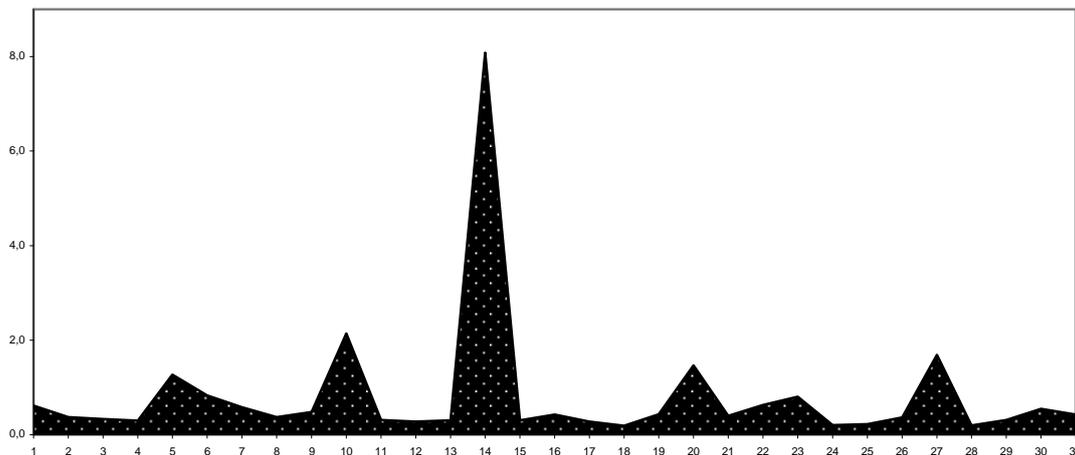


Fig. 27: La media de los valores absolutos de los coeficientes espaciales λ

4.3. Conclusiones

- ✓ Los modelos logarítmicos bivariados (usando precipitación acumulada en combinación con la humedad o temperatura) son los más idóneos para representar la influencia de las variables exógenas. Dado que los ajustes obtenidos fueron bastante satisfactorios se puede considerar que la hipótesis de que la dinámica de la malaria puede ser explicada a partir de la dinámica de variables meteorológicas fue verificada.
- ✓ El Niño es la medida de temperatura óptima para ser usada como regresor. Ahora esto podría deberse a que el estado Sucre es un estado costero, además de que la temperatura registrada es el promedio registrado en cierta cuadrante de superficie marina, mientras que la temperatura registrada en Guiria es un registro mas puntual. Habría que indagar en futuros estudios si esto sigue constatándose en regiones mas continentales
- ✓ El método de optimización siempre produce un incremento en los p-valores de los test de Ljung-Box y F, con lo que se hace menos probable que la hipótesis nula sea falsa. Pero solo en un tercio de los municipios este incremento fue notorio.
- ✓ Existe un correlación entre el incremento en los p-valores de los test de Ljung-Box y F y el incremento de la varianza.
- ✓ Los coeficientes espaciales tienen una distribución cuya forma se mantiene independiente del índice que se escoja. Esto constituye una evidencia a favor de que la hipótesis de que la dinámica de la malaria en cualquier localidad puede ser explicada a partir de la estructura espacio-temporal de la dinámica regional de la malaria.

5. Bibliografía

- Althizer, Sonia, *Seasonality and the dynamics of infectious diseases*, Ecology Letters, (2006) 9: 467–484
- Barrera R, Grillet ME, Rangel Y, Berti J y Ache A. *Estudio eco-epidemiológico de la reintroducción de la malaria en el nororiente de Venezuela mediante Sistemas de Información Geográfica y Sensores Remotos*. Boletín de la dirección de malariología y saneamiento ambiental. Vol. XXXVIII, N° 1, Enero - Julio, 1998.
- Basurko et al, *Influence of climate and river level on the incidence of malaria in Cacao, French Guiana*, Malaria Journal 2011, 10:26
- Baton L & Ranford-Cartwright L, *Spreading the seeds of million-murdering death: metamorphoses of malaria in the mosquito*. TRENDS in Parasitology Vol.21 No.12 December 2005,
- Breman J, *The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden*. Am J Trop Med Hyg 64 (1-2 Suppl): pp. 1-11. 2001.
- Briët, Olivier et al, *Malaria seasonality and rainfall seasonality in Sri Lanka are correlated in space*. Geospatial Health 2(2), 2008, pp. 183-190
- Brockwell P, David R, *Introduction to Time Series and Forecasting*, Springer-Verlag, 2002
- Chowell, Gerardo et al, *The spatial and temporal patterns of falciparum and vivax malaria in Perú: 1994–2006*, Malaria Journal 2009, 8:142
- Colón-González FJ, Lake IR & Bentham G, *Climate Variability and Dengue Fever in Warm and Humid Mexico*, Am. J. Trop. Med. Hyg., 84(5), 2011, pp. 757–763
- Del Barrio Castro T, *MODELOS DINAMICOS*, material de apoyo del curso de Econometría, de la Departament d'Economia Aplicada, Universitat de les Illes Balears <http://www.uib.es/depart/deaweb/webpersonal/hdeatbc0/archivos/Modelos_din.pdf>
- Grasso M. et al, *The Health Effects of Climate Change: A Survey of Recent Quantitative Research*, IEF Working Paper n. 27, January 2010.
- Grillet ME et al, *State transition detection in the spatio-temporal incidence of malaria*. Spatial and Spatio-temporal Epidemiology 1 251–259, 2010.
- Hoshen B and Morse A, *A weather-driven model of malaria transmission*. Malaria Journal 2004, 3:32, 2004.
- Koop G, *Bayesian Econometrics*, Wiley, 2003
- Le Sage J. *Applied Econometrics using MATLAB*, CRC Press/Taylor & Francis Group, 1999
- Paaijmans K et al, *Influence of climate on malaria transmission depends on daily temperature variation*. PNAS, vol. 107, n° 34 15135–15139. August 24, 2010.
- Parham PE & Michael E, *Modeling the Effects of Weather and Climate Change on Malaria Transmission*. Environmental Health Perspectives, vol 118 n° 5. Mayo 2010
- Rogers J et al, *Satellite imagery in the study and forecast of malaria*. NATURE VOL 415, 2002.
- Sáez-Sáez V. et al, *Evaluación semanal de la relación malaria, precipitación y temperatura del aire en la Península de Paria*, estado Sucre, Venezuela Bol Mal Salud Amb v.47 n.2. dic. 2007.
- Stressman G, *Beyond temperature and precipitation: Ecological risk factors that modify malaria transmission*. Acta Tropica 116 167–172. 2010
- Vilar Fernández JM, *Modelos Estadísticos Aplicados*, Publicaciones de la UDC, monografía 101. Segunda Edición. Junio, 2006. ISBN: 84-9749-196-3.
- Mapa de áreas de riesgo en Venezuela tomada de <http://www.sefar.gob.ve/index_1/malaria/Malaria%20en%20Venezuela.pdf>

Anexo: scripts del algoritmo de optimización

```

load NINO.mat; RN = RANGO(1,1);
load CLIMAGUIRIA.mat; RG = RANGO(1,1); aj2 = (RG - RN)*52;
load IPSparrsucre.mat; RY = info(1); Taj = (RY - RG)*52;
W = zeros(312,size(IPS,2));

for i = 1:size(IPS,2)
    stats = distLregrRES(log(IPS(:,i) + 1),[log(acumX(PP(:,3),10) +
1),log(NINO34((aj2+1):(aj2+size(TEMP,1)),3))],[52,52],[105,416],Taj);
    W(:,i) = stats.r; % residuo  $\{w_{k,t}\}$ 
end
Y = log(IPS(105:416,:) + 1); % transformada logarítmica del IPS  $\{y_t\}$ 
%%
p = 3; %orden del autorregresivo
LBumbral = 0.05; Fumbral = 0.05;
Tn = size(W,1); nloc = size(W,2); n = Tn - p;
v1 = 0.1; v2 = 0.2; %v1 y v2: varianzas del kernel de transicion
nit = 700; tb = 201; %nit: numero de iteracion; tb: iteraciones
quemadas
aPot = 3; %cociente del la funcion ALPHA
LBregistro = zeros(nit,nloc); Fregistro = zeros(nit,nloc); Vregistro =
zeros(nit,nloc);
%%
% inicializacion
PHIini = zeros(nloc,p); LAMBDAini = zeros(nloc,nloc);
PHImcmc = zeros(nloc,p); LAMBDAmcmc = zeros(nloc,nloc);
INDICES = zeros(nloc,6);
for loc = 1:nloc
    phi = yulewalker(W(:,loc),p); % estimando coeficientes phi
    % contruccion de la matriz V
    Xframe = Y; Xframe(:,loc) = W(:,loc); V = zeros(Tn-p,nloc);
    for i = 1:p
        V = V + phi(i)*Xframe((p+1-i):(Tn-i),:);
    end
    lambda = pinv(V)*W(p+1:Tn,loc); % estimando coeficientes lambda
    PHIini(loc,:) = phi; LAMBDAini(loc,:) = lambda;
    %indices iniciales
    z = W(p+1:Tn,loc) - V*lambda;
    Vopt = var(z);
    [~,LBopt] = lbqtest(z);
    Fopt = FDLtest(z,Xframe,p);
    INDICES(loc,1) = Vopt; INDICES(loc,3) = LBopt; INDICES(loc,5) =
Fopt;
    %cadena MCMC
    PHI = zeros(nit,p); LAMBDA = zeros(nit,nloc);
    for m = 1:nit
        %aplicar kernel de transicion
        [phi2,lambda2] = transkernel(phi,lambda,v1,v2);
        z = zCalc(W,Xframe,phi2,lambda2,loc); % recalculer z
        [~,LBult] = lbqtest(z); %aplicar test de Ljung-Box
        if LBult > LBumbral %2.a
            Fult = FDLtest(z,Xframe,p);
            if Fult > Fumbral %3.a
                Vult = var(z); alpha = min((Vopt/Vult)^aPot,1); u =
rand(1);
                if u <= alpha
                    phi = phi2; lambda = lambda2; Vopt = Vult;
                end
            end
        end
    end
end

```

```

        else
            if Fult > Fopt %3.b
                Vult = var(z); alpha = min((Vopt/Vult)^aPot,1); u
= rand(1);
                if u <= alpha
                    phi = phi2; lambda = lambda2; Vopt = Vult;
Fopt = Fult;
                end
            end
        else
            if LBult > LBopt %2.b
                Fult = FDLtest(z,Xframe,p);
                if Fult > Fumbral %3.a
                    Vult = var(z); alpha = min((Vopt/Vult)^aPot,1); u
= rand(1);
                    if u <= alpha
                        phi = phi2; lambda = lambda2; Vopt = Vult;
LBopt = LBult;
                    end
                else
                    if Fult > Fopt %3.b
                        Vult = var(z); alpha =
min((Vopt/Vult)^aPot,1); u = rand(1);
                        if u <= alpha
                            phi = phi2; lambda = lambda2; Vopt = Vult;
LBopt = LBult; Fopt = Fult;
                        end
                    end
                end
            end
            end
            PHI(m,:) = phi'; LAMBDA(m,:) = lambda';
            LBregistro(m,loc) = LBult; Fregistro(m,loc) = Fult;
Vregistro(m,loc) = Vult;
            end
            PHImcmc(loc,:) = sum(PHI(tb:nit,:))/(nit-tb+1);
            LAMBDAmcmc(loc,:) = sum(LAMBDA(tb:nit,:))/(nit-tb+1);
            %%
            %Indices MCMC
            for i = 1:p
                V = V + phi(i)*Xframe((p+1-i):(Tn-i),:);
            end
            z = W(p+1:Tn,loc) - V*lambda; z1 = z;
            INDICES(loc,2) = var(z); [~,INDICES(loc,4)] = lbqtest(z);
            INDICES(loc,6) = FDLtest(z,Xframe,p);
        end
        save('MCMC.mat','INDICES','LAMBDAini','LAMBDAmcmc','PHIini','PHImcmc')

```

En **negrita** están señaladas aquellas funciones que fueron creadas por mí

Funciones no nativas de matlab

```
function r = distLregrRES(y,x,r,tlimy,Taj)
%distributed lag regression
%argumentos: y, x, r, tlimy, tlimx
%y: variable respuesta
%x: variable(s) regresora(s)
%r: orden(es) del modelo DL
%tlimy: rango de y
%Taj: ajuste de tiempo de x respecto y
%salida:
%   r: Residual
n = tlimy(2) - tlimy(1) + 1; k = sum(r);
X = zeros(n,k); B = size(r,2); sk = 0;
tlimx = tlimy + Taj;
for i = 1:B
    xi = x(:,i);
    for j = 1:r(i)
        sk = sk + 1;
        X(:,sk) = xi(tlimx(1)-j:tlimx(2)-j);
    end
end
r = regstats(y(tlimy(1):tlimy(2)),X,'linear',{'r'});
```

```
function Xa = acumX(x,k)
%argumentos: x, k
n = size(x,1); Xa =zeros(size(x));
for t = k+1:n
    for j = 0:k
        Xa(t) = Xa(t) + x(t-j);
    end
end
end
```

```
function phi = yulewalker(x,p)
gamma = ACV(x,size(x,1));
GAMMA = zeros(p);
for i = 1:p
    for j = 1:p
        GAMMA(i,j) = gamma(abs(i-j)+1);
    end
end
end
gammap = zeros(3,1);
for i =1:p
    gammap(i) = gamma(i+1);
end
end
%%
if det(GAMMA)~=0
    phi = GAMMA\gammap;
else
    phi = pinv(GAMMA)*gammap;
end
end
```

```
function k = ACV(x,lags)
v = xcov(x,lags);
k = v(lags+1:2*lags+1);
```

```
function Fpval = FDLtest(z,X,p)
```

```

%distributed lag regression
%argumentos: y, X
%y: variable respuesta
%x: variable(s) regresora(s)
%salida:
%   Fpval: F p-value
n = size(X,1);
stats = regstats(z,X(p+1:n,:), 'linear', {'fstat'});
Fpval = stats.fstat.pval;

function [phi2,lambda2] = transkernel(phi,lambda,v1,v2)
%input: phi,lambda,v1,v2
%output: phi2,lambda258
n1 = size(phi,1); n2 = size(lambda,1);
s = [v1*ones(n1,1);v2*ones(n2,1)];
adj = normrnd(0, 1, (n1 + n2), 1).*s;
phi2 = phi + adj(1:n1);
lambda2 = lambda + adj(n1+1:n1+n2);

function z = zCalc(W,Xframe,phi,lambda,loc)
% input: W,Xframe,phi,lambda,loc
% output: z
p = size(phi,1); Tn = size(Xframe,1); nloc = size(Xframe,2);
V = zeros(Tn-p,nloc); w = W(p+1:Tn,loc);
for i = 1:p
    V = V + phi(i)*Xframe((p+1-i):(Tn-i),:);
end
z = w - V*lambda;

```