UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

MODELO ESPACIAL PARA EL RIESGO DE LA OBESIDAD EN MÉXICO

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS

PRESENTA:
ANGEL DE JESUS GUTIÉRREZ PRIETO

DIRECTOR
DR CARLOS DÍAZ AVALOS IIMAS

CO DIRECTOR
DRA NANCY RAQUEL MEJÍA DOMÍNGUEZ IIMAS

CIUDAD DE MÉXICO 9 DE NOVIEMBRE 2020.

Índice general

Capítulo 1

INTRODUCCIÓN

La Organización Mundial de la Salud (OMS) señala que en el mundo existe una pandemia de enfermedades crónicas no transmisibles la cual provoca un aproximado de 37 millones de muertes prematuras cada año[1]. En la categoría de las enfermedades crónicas no transmisibles existen al menos una docena de condiciones crónicas cuyo denominador común es un factor de riesgo que comprende los siguientes hábitos: alimentación inadecuada, sedentarismo y consumo de bebidas de alto contenido calórico los cuales están muy asociados al sobrepeso y obesidad.

Un informe de la OMS mencionó que en el año 2016, más de 1900 millones de adultos de 18 o más años tenían sobrepeso, de los cuales, más de 650 millones eran obesos y además también se señala que la prevalencia de obesidad se ha casi triplicado entre los años 1975 y 2016 [2].

La OMS también señala que la causa del sobrepeso y la obesidad es el desequilibrio energético entre calorías consumidas y gastadas. Además también se menciona que a nivel mundial ha ocurrido lo siguiente:

1. Hay aumento en la ingesta de alimentos de alto contenido calórico que son ricos en grasa.

2. Hay un descenso en la actividad física debido a la naturaleza cada vez más sedentaria de muchas formas de trabajo, los nuevos modos de transporte y la creciente urbanización.

En consecuencia, el padecer sobrepeso u obesidad se traduce en tener un alto índice de masa corporal el cual es un importante factor de riesgo de enfermedades no transmisibles, como las siguientes:

1. Enfermedades cardiovasculares (principalmente las cardiopatías y los accidentes cerebrovasculares), que fueron la principal causa de muertes en

2012

2. Diabetes

3. Trastornos del aparato locomotor (en especial la osteoartritis, una enfermedad degenerativa de las articulaciones muy discapacitante), y algunos cánceres (endometrio, mama, ovarios, próstata, hígado, vesícula biliar, riñones y colon).

Además, existe un mayor riesgo de padecer alguna de las enfermedades anteriores conforme aumenta el IMC y en este sentido (y a pesar de las controversias acerca de la interpretabilidad del IMC en términos de ser saludable o no ser saludable) el IMC provee una manera rápida de evaluar el riesgo de padecer una enfermedad no transmisible.

Particularmente en México durante los últimos 20 años ha existido un incremento considerable en los casos de sobrepeso y obesidad a tal grado que México se sitúa entre los países con mayor tasa de prevalencia de obesidad y sobrepeso. Para el año 2011, México ocupó el segundo lugar en peso excesivo en adultos y el cuarto lugar en niños entre los países integrantes de la OCDE[3] y el quinto en población adulta acorde a otros estudios [4] y en el año 2010 se estimó que el 75 % de todas las muertes fueron causadas por enfermedades no transmisibles donde la obesidad y dietas no saludables fueron catalogados entre los seis principales factores de riesgo[5].

El problema ha sido de tal magnitud que las proyecciones del gasto de salud para 13 enfermedades relacionadas con la obesidad fueron de 880 millones de dólares y se estima que para el año 2020 el gasto ascenderá a mil millones de dólares para atender esas enfermedades[6].

Un informe de la OCDE del 2014 menciona que el gasto total en salud de México representó un 6.2 % del producto interno bruto (PIB) lo cual situó a México como uno de los países que invirtió menos en términos porcentuales en ese rubro y que además la esperanza de vida se ha incrementado mucho mas lento que para otros países miembros de la OCDE. Este mismo informe menciona también que el lento crecimiento en la esperanza de vida en México se debe a comportamientos dañinos relacionados con la salud así como malos hábitos de nutrición y tasas muy altas de obesidad, como también al incremento en las tasas de mortalidad debido a la diabetes y a la no reducción en la mortalidad por enfermedades cardiovasculares [7].

Informes más recientes de la OCDE corroboran las tendencias anteriores. El informe del 2017 de esta organización señala que la tasa de crecimiento de la

esperanza de vida de los mexicanos se ha desacelerado a partir de los primeros años de la década del 2000 y el factor de riesgo más señalado para México es la prevalencia en obesidad el cual lo hace acreedor al segundo lugar en cuanto se refiere a prevalencia en adultos (con un valor del 33 %) y el primer lugar en la proporción general de sobrepeso y obesidad (con 73 %) entre los países miembros. En contraste, se señaló que México posee las tasas más bajas de consumo diario de tabaco (7.6 % mientras que el promedio de la OCDE está en 18.4 %) y en consumo de alcohol (5.2 litros de alcohol puro anual per cápita mientras que el promedio de la OCDE de está en 9.0 litros).

Y a pesar que el acceso a la atención médica (cuantificado por el seguro de salud) ha mejorado en México durante los últimos años, el acceso a recursos médicos está debajo de la media de la OCDE. Aunado a esto, el gasto en salud promedio (ajustado por los costos locales) es de $1,080$ USD por persona lo que es cuatro veces menos que el promedio de la OCDE que está en $4,003$ USD para el 2017.

Las altas tasas de prevalencia de obesidad están causando un impacto negativo en la calidad de vida de los mexicanos y acortando la esperanza de vida. México también está situado entre los países de la OCDE donde el consumo de frutas y verduras es de los más bajos y una tasa de diabetes de las más altas dentro de esta organización además, el sobrepeso y la obesidad es un factor de riesgo para muchas enfermedades crónicas.

Una de las estrategias que México ha implementado para mitigar el sobrepeso es el impuesto a bebidas azucaradas y restringir el acceso a niños y adolescentes ya que como menciona la OMS hay evidencia de que los niños y adolescentes tienen menos capacidad de ajustar sus hábitos alimenticios pensando en las consecuencias de largo plazo cuando hay una posibilidad de satisfacer un gusto o un placer de corto plazo. Incluso la OMS menciona que la industria es consciente de esta tendencia por lo cual utiliza la publicidad y el mercadeo de manera indiscriminada hacia los niños y adolescentes, cuando no hay una regulación eficaz para protegerlos [1].

En este sentido, una de las estrategias para medir la prevalencia del sobrepeso y obesidad en México ha sido a través de la Encuesta Nacional de Salud y Nutrición (ENSANUT). Esta encuesta contiene información acerca del numero de personas sin y con sobrepeso u obesidad en México e incluso detalla el municipio donde vive la persona encuestada así como sus medidas antropométricas.

1.1. FUNDAMENTO DEL PROBLEMA

Como se mencionó antes, uno de los retos más grandes que enfrenta México en el área de enfermedades no transmisibles son padecimientos como diabetes, enfermedades cardiovasculares y del aparato locomotor. Estos padecimientos lideran la causa de muertes en nuestro país y lo que tienen en común es el factor de riesgo de sobrepeso y la obesidad. En este sentido la motivación de este trabajo es plantear un modelo estadístico que permita la estimación y análisis de datos de la prevalencia de la obesidad y que pueda ser utilizado como apoyo en el proceso de la toma de decisiones; principalmente, en lo relativo al diagnóstico, evaluación y planificación de programas de salud de esta condición a nivel nacional.

El objetivo general de este trabajo es desarrollar un modelo estadístico que tome en cuenta factores socio económicos, edad, sexo y la variación espacial para explicar la relación entre riesgo relativo de padecer obesidad y los factores mencionados. Otro objetivo es la imputación del riesgo en municipios donde no hubo datos disponibles así como la identificación de conglomerados de municipios que destaquen más por su riesgo relativo.

Los datos que se usan en este trabajo son los resultados de la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2012 la cual es una encuesta probabilística con esquema de de muestreo poliétapico y estratificado [8]. La ENSANUT 2012 obtuvo estos resultados:

1. Tasa de repuesta para hogar de la ENSANUT 2012: 87 %.
2. Total de hogares con entrevista completa efectiva: $50,528$ hogares.
3. Total de entrevistas individuales completas en los hogares con entrevista completa efectiva: $96,031$.
4. Número de entrevistas completas de usuarios ambulatorios de los servicios de salud: Mayor a $14,104$.

El objetivo general de la ENSANUT 2012 es cuantificar la frecuencia, distribución y tendencias de las condiciones de salud y nutrición de la población y sus determinantes, así como examinar la respuesta social organizada frente a los problemas de salud y nutrición, incluida la cobertura y calidad de los servicios en la materia y la cobertura específica de los programas prioritarios de prevención en salud en los ámbitos nacional, estatal, por zonas urbanas y rurales, y por estratos socioeconómicos[8].

En este sentido, la ENSANUT 2012 provee resultados de las entrevistas individuales en términos de variables antropométricas como peso, talla, presión arterial así como edad y género. Los pasos que se realizaron para limpiar, estructurar y organizar la ingesta de los datos para la construcción de modelo fueron:

1. Para cada persona entrevistada se obtuvo su índice de masa corporal (IMC) basado en su peso y talla.

2. Utilizando el IMC, la edad y género, se clasificó a cada persona de la encuesta en las siguientes categorías: sana, sobrepeso, obesidad.[1]

3. Finalmente se contó el número de personas que padecen obesidad agrupadas por edad, sexo y municipio.[2]

4. Se estudiaron indices socioeconómicos por municipio para estudiar las relaciones entre estos y el riesgo relativo de la obesidad por municipio. En el modelo final sólo considera Índice de desarrollo humano y tasa de muertes asociadas a diabetes por cada 1000 habitantes.

El modelo propuesto permitirá evaluar si hay una relación entre los indicadores socioeconómicos y el número de casos de obesidad a través del riesgo de padecer obesidad y en el caso de los municipios donde no hubo observaciones (ya sea por que no se completaron las entrevistas o que los registros fueron inválidos) el modelo proveerá un estimado del riesgo relativo para esos municipios.

La tesis está estructurada en la siguiente forma. El capítulo 2 es un resumen de la teoría básica utilizada en estadística espacial. Ahí se introducen los conceptos de mallas y campos markovianos. También se menciona métodos de estimación de parámetros vía cadenas de Markov en modelos que utilizan campos markovianos. El capítulo 3 describe los pasos en la modelación del riesgo relativo de padecer obesidad utilizando los datos de la ENSANUT 2012. El capítulo 4 es una discusión acerca del método y los resultados obtenidos en la fase de modelación. Este capítulo también incluye la interpretación del modelo y mapas de riesgo. Las conclusiones de este trabajo están en el capítulo 5.

[1]Este trabajo está enfocado es estimar el riesgo de padecer obesidad por lo que las personas que tienen sobrepeso se consideran sanas.

[2]Debido a que el muestreo fue a nivel estatal y no municipal se estimó el número de personas que representa el encuestado en su municipio y no a nivel estatal.

Capítulo 2

MARCO TEÓRICO

2.1. Estadística espacial sobre mallas

De manera general, el problema que concierne a la estadística espacial es cuando se tiene un conjunto de locaciones $\boldsymbol{D} \subset \mathbb{R}^d$ y asumimos que las observaciones $\boldsymbol{Z}(\boldsymbol{s})$ provienen de una variable aleatoria. Entonces, variando $\boldsymbol{s} \in \boldsymbol{D}$ se obtiene un proceso aleatorio

$$\{\boldsymbol{Z}(\boldsymbol{s}) : \boldsymbol{s} \in \boldsymbol{D}\}$$

donde una realización de este proceso se denota como $\{\boldsymbol{z}(\boldsymbol{s}) : \boldsymbol{s} \in \boldsymbol{D}\}$. Usualmente, el conjunto $\boldsymbol{D}$ se asume fijo aunque se puede asumir que tanto $\boldsymbol{D}$ como $\boldsymbol{Z}$ pueden variar de realización a realización [9]. Si $\boldsymbol{D}$ es una colección contable fija de puntos en $\mathbb{R}^d$ donde $\boldsymbol{D}$ induce a una gráfica en $\mathbb{R}^d$ entonces diremos que el proceso $\boldsymbol{Z}(\boldsymbol{s})$ está definido sobre una malla.

Una manera intuitiva de construir una malla es por ejemplo a través de las entidades federativas de un país, donde cada entidad es representada por un vértice y las aristas son colocadas en base a la noción de "vecindad". Cabe señalar que hay distintas maneras de establecer la noción de vecindad. Por ejemplo, se puede usar la distancia euclideana o la del "taxista"para establecer si dos entidades son vecinas o si simplemente podemos decir que dos entidades son vecinas si tienen un borde en común.

2.2. Campos aleatorios Markovianos

Una parte importante en el estudio de los modelos espaciales es el concepto de campos aleatorios markovianos. Para un conjunto (finito) de sitios podemos establecer un puente entre una ley de probabilidad condicionada y la función de verosimilitud a través de la teoría de campos aleatorios markovianos.

Sea S el conjunto de zonas (en este caso municipios) y sea $s_i \in S$ la i-ésima zona espacial para $i = 1, ..., n$ donde n es el número total de municipios y sea $z(s_i)$ el valor observado de la variable aleatoria en la zona s_i y para simplificar la escritura, diremos que i y k son vecinos cuando las locaciones espaciales s_i y s_k sean vecinos. En este sentido, la estructura de vecindad es importante en la modelación de fenómenos espaciales ya que es de interés hacer inferencia estadística usando leyes de probabilidad condicionadas a valores observados de los vecinos del lugar donde se hace la observación, así que uno de los conceptos importantes en el análisis de la teoría de campos aleatorios markovianos es el de *independencia condicional*.

Definición 1 *Dos variables aleatorias X y Y son condicionalmente independientes dado Z sí y sólo sí la distribución conjunta de X y Y dado Z se puede factorizar como el producto de la marginal de X y Y i.e:*

$$\pi(X, Y|Z) = \pi(X|Z)\pi(Y|Z)$$

Teorema 1 *Sean X, Y y Z variables aleatorias. Entonces X y Y son condicionalmente independientes dado Z sí y sólo sí $\pi(X, Y|Z) = f(X, Z)g(Y, Z)$ para algunas funciones f y g, y para todo Z tal que $\pi(Z) > 0$*

Más aún, el concepto de independencia condicional en los campos aleatorios markovianos es representado a través de gráficas no dirigidas.

Definición 2 *Una gráfica G no dirigida es una tupla (V, E), donde V es el conjunto de vertices de G y E el conjunto de aristas $\{i, j\}$.*

En las aplicaciones que corresponden a estadística espacial, el conjunto V de la definición anterior son las locaciones $V = S$ (y en este caso particular municipios) y el conjunto de aristas E se define en base a las vecindades (adyacencias) de cada locación s_i.

Definición 3 *Sean $\{s_i : 1, ..., n\}$ el conjunto de las n zonas en una malla en los cuales una variable aleatoria $\boldsymbol{Z}$ es observada. Sea $\zeta \equiv \{\boldsymbol{z} : \pi(\boldsymbol{z}) > 0\}$ y $\zeta_i \equiv \{z(s_i) : \pi(z(s_i)) > 0\}$ para $i = 1, ..., n$. Entonces decimos que la **condición de positividad** se cumple si $\zeta = \zeta_1 \times ... \times \zeta_n$. En el caso en que la variable aleatoria $\boldsymbol{Z}$ sea continua pediremos que $\zeta \equiv \{\boldsymbol{z} : f(\boldsymbol{z}) > 0\}$ y $\zeta_i \equiv \{z(s_i) : f_i(z(s_i)) > 0\}$ donde f denota a la función de densidad de $\boldsymbol{Z}$ y f_i a la densidad marginal respecto a $Z(s_i)$.*

La importancia de la condición de positividad radica en que si ésta se satisface, entonces la ley de probabilidad de cualquier campo aleatorio markoviano está unicamente determinado por sus leyes de probabilidad condicional

"locales"[10]. Es decir, si $\pi(z(s_i)|\{z(s_j)j \neq i, i = 1, ..., n\})$ denota a la ley condicional de $z(s_i)$ dado los valores observados en los sitios s_j con $i \neq j$ y si la factorización de Besag se cumple, entonces la ley conjunta $\pi(\boldsymbol{z})$ está bien definida.

Teorema 2 (Teorema de factorización de Besag) *Supóngase que las variables aleatorias $\{Z(s_i) : i = 1, ..., n\}$ tienen como función de masa de probabilidad a $\pi(\cdot)$ y sea ζ el soporte de $\pi(\cdot)$ tal que ζ satisface la condición de positividad. Entonces,*

$$\frac{\pi(\boldsymbol{z})}{\pi(\boldsymbol{y})} = \prod_{i=1}^{n} \frac{\pi\left[z(s_i)|z(s_1), ..., z(s_{i-1}), y(s_{i+1}), ..., y(s_n)\right]}{\pi\left[y(s_i)|z(s_1), ..., z(s_{i-1}), y(s_{i+1}), ..., y(s_n)\right]}, \quad \boldsymbol{y}, \boldsymbol{z} \in \zeta$$

donde $\boldsymbol{y}^t = (y(s_1), ..., y(s_n))$ y $\boldsymbol{z}^t = (y(s_1), ..., y(s_n))$ denotan a las posibles realizaciones de la variable aleatoria Z.

Definición 4 *El sitio s_k es vecino del sitio s_i si la distribución condicional de $Z(s_i)$ dado los otros valores de los sitios, depende funcionalmente de $z(s_k)$ para $k \neq i$. También definimos*

$$N_i \equiv \{k : s_k \text{ es vecino de } s_i\} \tag{2.1}$$

como el conjunto vecindad del sitio i.

En el caso cuando se haga referencia a un vector con diferentes valores observados z que pertenecen a un subconjunto de C de locaciones en S lo denotaremos como $\boldsymbol{z_C} = \{z(s_i) : s_i \in C \subseteq S\}$.

Definición 5 *Una pandilla se define como el conjunto de sitios que consiste de un sólo sitio o sitios en el que todos son vecinos entre si.*

Definición 6 *Cualquier medida de probabilidad cuya distribución condicional defina una estructura de vecindad $\{N_i : i = 1, ...n\}$ través de la expresión (2.1) se dice ser un* ***campo aleatorio markoviano****.*

La definición anterior es muy importante ya que nos permite construir un campo aleatorio markoviano a través de establecer una ley de probabilidad donde la distribución de $Z(s_i)$ condicionada a todos los valores restantes dependa únicamente de sus vecinos. Es decir, si $\pi(\boldsymbol{z}) > 0$ para todo $\boldsymbol{z}$ y $\pi(z(s_i)|z_{V-\{i\}}) = \pi(z(s_i)|z_{N_i})$ entonces $\boldsymbol{Z}$ es un campo aleatorio markoviano con respecto a la gráfica (V, E)[1].

[1]La notación $z_{V-\{i\}}$ denota a los valores observados de $\boldsymbol{Z}$ salvo en el vértice i y z_{N_i} denota a los valores observados de $\boldsymbol{Z}$ en los vecinos de i

Esencialmente, hay dos enfoques para establecer o especificar un campo aleatorio markoviano. El primero es en términos de las leyes condicionales $\pi(z(s_i)|z_{N_i})$ y el segundo en términos de la ley conjunta $\pi(\boldsymbol{z})$. El primer enfoque tiene la desventaja de que no hay un método obvio para establecer esa ley de probabilidad conjunta y el segundo enfoque tiene la desventaja de que puede estar sujeto a condicionales altamente restrictivas y no obvias[11]. Afortunadamente, existe un resultado teórico que garantiza la existencia y unicidad de las leyes de probabilidad condicionadas para $Z(s_i)$. Esto se logra a través de la función *neg-potencial*.

Definición 7 *El vector aleatorio* $\boldsymbol{Z} = (Z(s_1), ..., Z(s_n))$ *se dice ser un campo aleatorio de Gibbs con respecto a la gráfica* (V, E) *si y sólo sí las realizaciones de* $\boldsymbol{Z}$ *siguen una distribución de Gibbs.*

Una distribución de Gibbs tiene la siguiente forma:

$$\pi(\boldsymbol{z}) = \frac{\exp\left(-\frac{Q(\boldsymbol{z})}{T}\right)}{\sum_{\boldsymbol{y}\in\zeta}\exp\left(-\frac{Q(\boldsymbol{y})}{T}\right)} \tag{2.2}$$

donde $\sum_{\boldsymbol{y}\in\zeta}\exp\left(-\frac{Q(\boldsymbol{y})}{T}\right)$ es la constante de normalización también llamada *función de partición*, T es una constante que representa a la temperatura (se asumirá igual 1) y $Q(\boldsymbol{z})$ es la *función de energía* en el contexto de mecánica estadística.

Sea $\zeta \equiv \{\boldsymbol{z} : \pi(\boldsymbol{z}) > 0\}$ con $\boldsymbol{z} \in \mathbb{R}^n$ y supongamos que $\mathbf{0} \in \zeta$ es decir, es posible observar 0 en todas zonas por lo que $\pi(\mathbf{0}) > 0$. Definamos

$$Q(\boldsymbol{z}) \equiv \log\left[\frac{\pi(\boldsymbol{z})}{\pi(\mathbf{0})}\right], \quad \boldsymbol{z} \in \zeta \tag{2.3}$$

entonces la función *neg-potencial* o de *energía* satisface la siguientes propiedades:

Lema 1 *Sea* Q *como se definió en la expresión (2.3), entonces* Q *cumple las siguientes propiedades:*

1. $$\frac{\pi\left[z(s_i)|\{z(s_j) : j \neq i\}\right]}{\pi\left[0(s_i)|\{z(s_j) : j \neq i\}\right]} = \frac{\pi(\boldsymbol{z})}{\pi(\boldsymbol{z}_i)} = \exp(Q(\boldsymbol{z})) - \exp(Q(\boldsymbol{z}_i))$$

   donde $0(s_i)$ *denota al evento* $Z(s_i) = 0$ *y* $\boldsymbol{z}_i \equiv (z(s_1), ..., z(s_{i-1}), 0, z(s_{i+1}), ..., z(s_n))^t$

2. Q se puede expandir de manera única sobre ζ como:

$$\begin{aligned} Q(\boldsymbol{z}) = &\sum_{1\leqslant i\leqslant n} z(s_i)G(z(s_i)) + \sum_{1\leqslant i<j\leqslant n}\sum z(s_i)z(s_j)G_{ij}(z(s_i),z(s_j)) \\ &+ \sum\sum_{1\leqslant i<j<k\leqslant n}\sum z(s_i)z(s_j)z(s_k)G_{ijk}(z(s_i),z(s_j),z(s_k)) \\ &+ z(s_1)\cdots z(s_n)G_{1\cdots n}(z(s_1),...,z(s_n)), \quad \textit{con } \boldsymbol{z}\in\zeta. \end{aligned} \tag{2.4}$$

Como se menciona en lema 1, la expansión (2.4) es única pero las funciones $\{G_{ij\cdots}\}$ no están definidas de manera única. Para garantizar la unicidad basta hacer $G_{ij\cdots}\equiv 0$ siempre que alguno de los argumentos $z(s_i)$ o $z_(s_j)$ de la función $G_{ij\cdots}$ sea igual 0 [9].
Además, derivado de propiedades del Lema 1 y considerando el evento

$$\{z(s_j)=0 \text{ para } j\neq i\}$$

se tiene que

$$\frac{\pi\left[z(s_i)|\{z(s_j)=0:j\neq i\}\right]}{\pi\left[0(s_i)|\{z(s_j)=0:j\neq i\}\right]} = \exp(Q(z)) = \exp\left[z(s_i)G_i(z(s_i))\right]$$

por lo que

$$z(s_i)G_i(z(s_i)) = \log\left[\frac{\pi\left[z(s_i)|\{z(s_j)=0:j\neq i\}\right]}{\pi\left[0(s_i)|\{z(s_j)=0:j\neq i\}\right]}\right] \tag{2.5}$$

es decir, los términos que corresponden al efecto de $z(s_i)$ están asociados a los cocientes de las distribuciones condicionales de $z(s_i)$ y $0(s_i)$ dado los valores vecinos s_j según las expresiones (2.5) y (2.2). Además, como $\pi(\boldsymbol{z})\propto\exp(Q(\boldsymbol{z}))$ entonces la ley conjunta de Z puede ser caracterizada por Q en términos de los soportes de las distribuciones marginales de Z así como de sus interacciones a pares acorde a (2.2).

Aunque la función Q puede ser caracterizada, queda abierto el encontrar la constante de proporcionalidad de la expresión (2.2). Esta tarea no es fácil ya que la función de verosimilitud puede depender de parámetros adicionales. Aún así, el siguiente teorema ayuda a caracterizar la forma que debe tener la función $Q(\cdot)$ tomando en cuenta la estructura de vecindad del campo markoviano.

Teorema 3 (Hammersley - Clifford) *Supóngase que la variable aleatoria $\boldsymbol{Z}$ se distribuye acorde a un campo aleatorio Markoviano sobre ζ y que satisface la condición de positividad. Entonces la función neg-potencial $Q(\cdot)$ de la expresión (2.4) debe satisfacer la siguiente propiedad:*

si las zonas $i, j, ..., s$ *no forman una pandilla, entonces* $G_{ij\cdots s}(\cdot) \equiv 0$,

donde la pandilla esta definida por la estructura de las vecindades $\{N_i : 1, ..., n\}$.

La importancia del Teorema de Hammersley-Clifford radica en que establece la equivalencia entre la propiedad local y global de Markov. Es decir, mientras un campo aleatorio markoviano es caracterizado por la propiedad *local de Markov*, un campo aleatorio de Gibbs está caracterizado por la propiedad *global*.

En la práctica, el valor de este teorema es que provee una manera de especificar la ley conjunta de $\boldsymbol{Z}$, es decir, la densidad o función de masa de probabilidad $\pi(\boldsymbol{z})$ puede ser factorizada sobre *pandillas* o equivalentemente sub-gráficas completas de la gráfica (V, E).

Una vez que la estructura de vecindad está definida, queda abierto la especificación de las leyes condicionales de $\pi(z(s_i)|\{z(s_j) : j \neq i\})$. Para el caso en que estas leyes condicionales pertenecen a la familia exponencial ha sido estudiado de manera amplia y en particular para aplicaciones que tienen que ver con el análisis espacial del riesgo de enfermedades, la elección natural es utilizar variables aleatorias Poisson y modelar el riesgo a través de tasa de eventos. De manera general, si la ley de $z(s_i)$ condicionado a sus vecinos es miembro de la familia exponencial, entonces ésta se puede escribir de la siguiente forma:

$$\pi\left[z(s_i)|\{z(s_j) : j \neq i\}\right] = \exp\Big[A_i(\{z(s_j) : j \neq i\})B_i(z(s_i)) + C_i(z(s_i)) + D_i(\{z(s_j) : j \neq i\})\Big], \text{ para } i = 1, ..., n \tag{2.6}$$

donde $A_i(\cdot)$ y $D_i(\cdot)$ son funciones que dependen los lo valores vecinos del sitio s_i. La representación (2.6) es conveniente ya que el siguiente teorema da indicios de la estructura debe tener la función A_i y además lo liga con la *función neg-potencial* Q siempre que se trabaje con distribuciones que pertenecen a la familia exponencial.

Teorema 4 *(Besag,1974) Sea A un conjunto de indices y sea Z una variable aleatoria cuya ley de probabilidad pertenece a la familia exponencial. Asúmase dependencia a pares entre sitios (i.e si A contiene tres o más índices diferentes entonces* $G_A \equiv 0$ *en la terminología de la expresión (2.4)). Entonces,*

$$A_i(\{z(s_j) : j \neq i\}) = \alpha_i + \sum_{j=1}^{n} \theta_{ij} B_j(z(s_j)), \text{ para } i = 1, ..., n \tag{2.7}$$

donde $\theta ij = \theta_{ji}$, $\theta_{ii} = 0$ *y* $\theta_{ik} = 0$ *si* $k \notin N_i$.

De manera más concreta, una ley de probabilidad que se establece a partir de la función de energía Q que sólo toma en cuenta las dependencias a pares derivará en una subclase de campos aleatorios markovianos denominados automodelos. Bajo este supuesto, la función Q tiene esta forma:

$$Q(\boldsymbol{z}) = \sum_{i=1}^{n} z(s_i)G_i(z(s_i)) + \sum_{i<j}\sum z(s_i)z(s_j)\theta_{ij}$$

2.3. Campos Aleatorios Markovianos Gaussianos

Distribución normal multivariada y propiedades básicas

La densidad de una variable aleatoria normal multivariada $\boldsymbol{x} = (x_1, ..., x_n)^t$ con media $\boldsymbol{\mu} \in \mathbb{R}^n$ y matriz de varianzas y covarianzas simétrica definida positiva $\boldsymbol{\Sigma}_{n\times n}$ es:

$$\pi(\boldsymbol{x}) = (2\pi)^{(n/2)}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad \boldsymbol{x} \in \mathbb{R}^n$$

donde $\mathbb{E}(x_i) = \mu_i$, $\Sigma_{ij} = Cov(x_i, x_j)$ y $\Sigma_{ii} = Var(x_i) > 0$. En este trabajo se optará por usar la matriz de precisión $\boldsymbol{Q}$, la cual es la inversa de la matriz $\boldsymbol{\Sigma}$ i.e $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$

Teorema 5 *Sea $\boldsymbol{x}$ una variable aleatoria normal con media $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{Q}$ semi definida positiva. Entonces para $i \neq j$ x_i y x_j son condicionalmente independientes dado $\boldsymbol{x}_{-ij}$ sí y sólo sí $Q_{ij} = 0$*

en el teorema anterior $\boldsymbol{x}_{-ij} = \boldsymbol{x}_{-\{i,j\}}$, es decir estamos condicionando sobre el vector que no considera la arista $\{i, j\} \in E$ lo cual implica que se puede deducir la estructura de la gráfica G mediante ver las entradas que son diferentes de 0 en la matriz $\boldsymbol{Q}$.

Definición 8 *Un vector aleatorio $\boldsymbol{x} = (x_1, ..., x_n)^T \in \mathbb{R}^n$ es un **campo aleatorio markoviano gausssiano** con respecto a una gráfica etiquetada $G = (V, E)$ con media $\boldsymbol{\mu}$ y matriz de precisión semi definida positiva $\boldsymbol{Q}$ sí y sólo sí su densidad tiene la forma*

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{Q}|^{1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

y $Q_{ij} \neq 0$ sí y sólo sí $\{i, j\} \in E$ para todo $i \neq j$.

Teorema 6 *Sea $\boldsymbol{x}$ un campo aleatorio markoviano gaussiano respecto a una gráfica $G = (V, E)$ con media $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{Q}$ simétrica definida positiva , entonces*

$$\begin{aligned} \mathbb{E}(x_i|\boldsymbol{x}_{-i}) &= \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ii}(x_j - \mu_j), \\ Prec(x_i|\boldsymbol{x}_{-i}) &= Q_{ii} \\ Corr(x_i, x_j|\boldsymbol{x}_{-ij}) &= -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, i \neq j \end{aligned} \tag{2.8}$$

Teorema 7 *Sea $\boldsymbol{x}$ un campo aleatorio markoviano gaussiano con respecto a la gráfica $G = (V, E)$ con media $\boldsymbol{\mu}$ y matriz de precisión $\boldsymbol{Q}$ semi definida positiva. Sea $A \subset V$ y $B = V - A$ donde $A, B \neq \emptyset$. La distribución condicional de $\boldsymbol{x}_a|\boldsymbol{x}_b$ es un campo aleatorio markoviano gaussiano respecto a la sub gráfica G^A no media $\boldsymbol{\mu}_{A|B}$ y matriz de precisión $\boldsymbol{Q}_{A|B}$ semi definida positiva donde*

$$\begin{aligned} \boldsymbol{\mu}_{A|B} &= \boldsymbol{\mu}_A - \boldsymbol{Q}_{AA}^{-1}\boldsymbol{Q}_{AB}(\boldsymbol{x}_B - \boldsymbol{\mu}_B), \\ \boldsymbol{Q}_{A|B} &= \boldsymbol{Q}_{AA} \end{aligned} \tag{2.9}$$

De manera alterna se puede especificar un campo aleatorio markoviano gaussiano a través de sus condicionales completas $\{\pi(x_i|\boldsymbol{x}_{-i})\}$. En la literatura a estos modelos se les conoce como autorregresiones condicionales (*conditional autoregressions*) y fueron desarrollados por Besag.

Definición 9 *Un campo aleatorio markoviano gaussiano $\boldsymbol{x}$ con respecto a una gráfica G con parámetros canónicos $\boldsymbol{b}$ y matriz simétrica definida positiva $\boldsymbol{Q}$ tiene densidad*

$$\pi(\boldsymbol{x}) \propto \exp(-\frac{1}{2}\boldsymbol{x}^t\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^t\boldsymbol{x})$$

donde la matriz de precisión es $\boldsymbol{Q}$ y la media es $\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{b}$. Esta parametrización canónica será denotada como $\boldsymbol{x} \sim N_C(\boldsymbol{b}, \boldsymbol{Q})$ y su relación con la distribución normal es $\boldsymbol{x} \sim N_C(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q}) = N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$

Teorema 8 *Dado n distribuciones condicionales completas con media condicional y precisión*

$$\begin{aligned} \mathbb{E}(x_i|\boldsymbol{x}_i) &= \mu_i - \sum_{j:j\sim i} \beta_{ij}(x_j - \mu_j) \\ Prec(x_i|\boldsymbol{x}_{-i}) &= \kappa_i > 0 \end{aligned} \tag{2.10}$$

para $i = 1, ..., n$ *y algunos coeficientes* $\{\beta_{ij}, i \neq j\}$, *entonces* $\boldsymbol{x}$ *es un campo aleatorio markoviano gaussiano con respecto a la gráfica* $G = (V, E)$ *con media* $\boldsymbol{\mu}$ *y matriz de precisión* $\boldsymbol{Q} = (Q_{ij})$, *donde*

$$Q_{ij} = \begin{cases} \kappa_i \beta_{ij}, & i \neq j \\ \kappa_i, & i = j \end{cases}$$

con $\kappa_i \beta_{ij} = \kappa_j \beta_{ji}$ *para* $i \neq j$ *y* (Q) *matriz.*

El teorema anterior es importante ya que formaliza la existencia y caracteriza al campo aleatorio markoviano gaussiano basado en las distribuciones condicionales completas de x_i. La prueba del teorema toma como herramienta al Lema de Brook:

Lema 2 *(Brook) Sea* $\pi(\boldsymbol{x})$ *la densidad de una variable aleatoria* $\boldsymbol{x}$ *que toma valores en* $\mathbb{R}^n$ *y sea* $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n : \pi(\boldsymbol{x}) > 0\}$. *Sea* $\boldsymbol{x}$ *y* $\boldsymbol{x}^{'} \in \mathbb{R}^n$, *entonces*

$$\begin{aligned} \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x}^{'})} &= \prod_{i=1}^{n} \frac{\pi(x_i | x_1, ..., x_{i-1}, x_{i+1}^{'}, ..., x_n^{'})}{\pi(x_i^{'} | x_1, ..., x_{i-1}, x_{i+1}^{'}, ..., x_n^{'})} \\ &= \prod_{i=1}^{n} \frac{\pi(x_i | x_1^{'}, ..., x_{i-1}^{'}, x_{i+1}, ..., x_n)}{\pi(x_i^{'} | x_1^{'}, ..., x_{i-1}^{'}, x_{i+1}, ..., x_n)} \end{aligned} \tag{2.11}$$

Demostración del Lema de Brook: Por probabilidad condicional se tiene

$$\frac{\pi(x_n | x_1, ..., x_{n-1}) \pi(x_1, ..., x_n)}{\pi(x_n^{'} | x_1, ..., x_{n-1}) \pi(x_1, ..., x_n^{'})} = \frac{\pi(x_1, ..., x_n)}{\pi(x_1, ..., x_n^{'})}$$

lo cual implica

$$\pi(x_1, ..., x_n) = \frac{\pi(x_n | x_1, ..., x_{n-1})}{\pi(x_n^{'} | x_1, ..., x_{n-1})} \pi(x_1, ..., x_n^{'})$$

y aplicando de manera inductiva ahora sobre la entrada x_{n-1}, se tiene

$$\pi(x_1, ..., x_n) = \frac{\pi(x_n | x_1, ..., x_{n-1})}{\pi(x_n^{'} | x_1, ..., x_{n-1})} \frac{\pi(x_{n-1} | x_1, ..., x_{n-2}, x_n^{'})}{\pi(x_{n-1}^{'} | x_1, ..., x_{n-2}, x_n^{'})} \pi(x_1, ..., x_{n-2}, x_{n-1}^{'}, x_n^{'})$$

por lo que

$$\frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x}^{'})} = \prod_{i=1}^{n} \frac{\pi(x_i | x_1, ..., x_{i-1}, x_{i+1}^{'}, ..., x_n^{'})}{\pi(x_i^{'} | x_1, ..., x_{i-1}, x_{i+1}^{'}, ..., x_n^{'})}$$

La segunda igualdad del Lema de Brook se prueba de manera similar.

2.4. Modelo Poisson y campos aleatorios markovianos

A pesar de que por definición un campo aleatorio markoviano gaussiano sigue una distribución normal es posible combinar modelos de respuesta no normal con variables de entrada modeladas con campos aleatorios markovianos. Por ejemplo, si una observación y se distribuye Poisson con media λ y sea $\eta = log(\lambda)$ donde se establece como distribución a priori $\eta \sim N(\mu, k^{-1})$ entonces la distribución posterior para η dado la observación y es

$$\begin{aligned} \pi(\eta|y) \propto \pi(y|\eta)\pi(\eta) &= \exp(-e^{\eta}) \exp(\eta)^{y} \exp(-\frac{k}{2}(\eta-\mu)^2) \\ &= \exp\left(-\frac{k}{2}(\eta-\mu)^2 + y\eta - \exp(\eta)\right) \end{aligned} \tag{2.12}$$

donde el objetivo es hacer inferencias sobre η usando (2.12) mediante una muestra de la distribución posterior de η. Una manera de abordar este problema es usando la aproximación de segundo orden de Taylor para $f(\eta) = \exp\left(-\frac{k}{2}(\eta-\mu)^2 + y\eta - \exp(\eta)\right)$ alrededor de algún valor razonable η_0. Es decir

$$\begin{aligned} f(\eta) \approx & f(\eta_0) + f^{'}(\eta_0)(\eta-\eta_0) + \frac{1}{2}f^{''}(\eta_0)(\eta-\eta_0)^2 \\ = & a + b\eta - \frac{1}{2}c\eta^2 \end{aligned} \tag{2.13}$$

con $a = f(\eta_0) - f^{'}(\eta_0)\eta_0 + \frac{1}{2}f^{''}(\eta_0)(\eta_0^2)$, $b = f^{'}(\eta_0) - f^{''}(\eta_0)\eta_0$ y $c = -\frac{1}{2}f^{''}(\eta_0)$.

La idea es aproximar a $\pi(y|\eta)$ a través del algoritmo de Metropolis-Hastings utilizando la aproximación de Taylor de $f(\eta)$, y dado que a es un constante, la densidad propuesta (sin normalizar) para $\pi(\eta|y)$ es:

$$\tilde{\pi}(\eta|y) \propto \exp\left(-\frac{1}{2}c\eta^2 + b\eta\right) \tag{2.14}$$

donde claramente esta densidad está parametrizada acorde a la forma canónica de una variable aleatoria normal $N_C(b,c) = N(b/c, c)$, es decir, se utilizarán muestras de distribuciones normales para aproximar la distribución $\pi(\eta|y)$. Cabe señalar que b y c dependen del valor η_0 y este valor se irá actualizando con las muestras que sean aceptadas durante la ejecución del algoritmo de Metropolis-Hastings.
Recordando que el objetivo es generar muestras de $\pi(\eta|y)$ a través de $\tilde{\pi}(\eta|y)$ el algoritmo de Metropolis-Hastings aceptará el valor η^* generado de una variable

aleatoria normal con media $\mu(\eta_0) = b/c$ y precisión $k(\eta_0)$ y lo aceptamos con probabilidad

$$\alpha = \text{mín}\left\{1, \frac{\pi(\eta^*|y)}{\pi(\eta_0|y)}\frac{\tilde{\pi}(\eta_0|\eta^*)}{\tilde{\pi}(\eta^*|\eta_0)}\right\}$$

y además el valor η_0 será actualizado por candidatos η^* que sean aceptados.

Análogamente, si $y_1, ..., y_n$ son observaciones condicionalmente independientes de una distribución distinta a la normal tal que y_i es una observación indirecta de x_i y además $\boldsymbol{x}$ es un un campo aleatorio markoviano gaussiano con matriz de precisión $\boldsymbol{Q}$ y media $\boldsymbol{\mu}$ entonces la distribución posterior queda de esta manera:

$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu}) + \sum_{i=1}^{n}\log(\pi(y_i|x_i))\right) \tag{2.15}$$

donde la ecuación (2.15) tiene incluido el logaritmo de la verosimilitud en lugar de la verosimilitud, ésto debido a que puede ayudar a incrementar la tasa de aceptación en el algoritmo de Metropolis-Hastings. Además tomando en cuenta la expansión de Taylor de la ecuación (2.13) para $\pi(y_i|x_i)$ al rededor de un punto $\boldsymbol{\mu_0}$, se tiene que

$$\begin{aligned}\tilde{\pi}(\boldsymbol{x}|\boldsymbol{y}) &\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^t\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{\mu}^t\boldsymbol{Q}\boldsymbol{x} + \sum_i(a_i + b_ix_i - \frac{1}{2}c_ix_i^2)\right) \\ &\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^t(\boldsymbol{Q} + diag(\boldsymbol{c}))\boldsymbol{x} + (\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{b})^t\boldsymbol{x}\right)\end{aligned} \tag{2.16}$$

donde $diag(\boldsymbol{c})$ es una matriz diagonal cuyos elementos son $c_1, ..., c_n$. y en el caso en que sean negativos se tomarán como 0 en $diag(\boldsymbol{c})$. Entoces, la parametrización canónica de $\tilde{\pi}(\boldsymbol{x}|\boldsymbol{y})$ es $N_C(\boldsymbol{Q\mu + b, Q + diag(c)})$. Además, de manera similar al ejemplo de la ecuación (2.14, se tiene que $\tilde{\pi}(\boldsymbol{x}|\boldsymbol{y})$, $\boldsymbol{b}$ y $\boldsymbol{c}$ depende de $\boldsymbol{\mu_0}$. Más aun, la distribución propuesta en (2.16) puede ser aproximada por Taylor alrededor de su media $\boldsymbol{\mu_1}$ y sucesivamente la aproximación pude ser mejorada mediante la expansión de Taylor alrededor de la media de $N_C(\boldsymbol{Q\mu + b, Q + diag(c)})$. Por otro lado, recordando que $\boldsymbol{x}$ es un campo aleatorio markoviano gaussiano y viendo que en la ecuación (2.16) las y_i dependen sólo de x_i entonces se tiene que $N_C(\boldsymbol{Q\mu + b, Q + diag(c)})$ hereda la propiedad de Markov y cuando se incluyan covariables z_i se tendrá que y_i dependerá de $z_i^t\boldsymbol{x}$ y entonces la propiedad de Markov puede o no ser heredada dependiendo de las covariables z_i [12].

2.4.1. Modelación conjunta del riesgo relativo de enfermedades

Cuando se tienen datos que son conteos de alguna enfermedad por área, el modelo más natural para trabajar con ellos es el modelo Poisson:

$$y_i \sim Poi(c_i e^{x_i})$$

donde c_i es el número de casos esperados en el área i, x_i es el logaritmo del riesgo relativo y las observaciones $y_1, ..., y_n$ se asumen condicionalmente independientes. El modelo propuesto por Besag, York y Mollie (BYM) [13] en presencia de covariables para explicar el logaritmo del riesgo relativo es planteado de esta forma:

$$x_i = z_i^t \boldsymbol{\beta} + u_i + v_i$$

donde $\boldsymbol{\beta}^t = (\beta_1, ..., \beta_p)$ denota al vector de parámetros asociado a las covariables.

El parámetro u_i es el encargado modelar el efecto espacial y la distribución condicional de u_i dado sus vecinos $\boldsymbol{u}_{-i} = (u_1, ..., u_{i-1}, u_{i+1}, ..., u_n)$ es:

$$u_i|\boldsymbol{u}_{-i}, k \sim N\left(\frac{\sum_{j=1}^n w_{ij}u_j}{\sum_{j=1}^n w_{ij}}, \frac{k}{\sum_{j=1}^n w_{ij}}\right), \text{ con } k \sim \text{Gama } (a, b) \qquad (2.17)$$

donde la estructura de autocorrelación o dependencia espacial es establecida por la matriz (simétrica) de adyacencias cuyas entradas son w_{ij} con $w_{ij} = 1$ si la locación s_i es vecina de s_j (lo denotaremos como $i \sim j$) y $w_{ij} = 0$ en otro caso. Además, en la expresión (2.17) se observa que la esperanza condicional del $i-$ésimo efecto asociado a la estructura espacial es el promedio de los efectos espaciales de los vecinos de la localidad i y la varianza condicional es inversamente proporcional al número de vecinos de i.

Por otro lado, en el caso de los parámetros v_i la distribución a priori que tomaremos es

$$v_i \sim N(0, \lambda)$$

donde $\lambda \sim$ Gamma Inversa(a,b), y los valores de a y b están preestablecidos.

Una limitante del modelo BYM es que x_i depende de dos efectos aleatorios (además del efecto asociado a las covariables) por lo que identificar a los parámetros u_i y v_i de manera separada es complicado, sin embargo la suma de éstos si es identificable [14].

2.4.2. Planteamiento del Modelo

En esta sección describimos el problema y la notación para el ajuste del modelo. En nuestro caso particular estaremos trabajando con el mapa de México dividido a nivel de municipios. El conjunto de municipios es $\{s_i, i = 1, ..., n\}$, es decir s_i denota al municipio i aunque para simplificar la notación diremos municipio i cuando hagamos referencia a s_i y el total de municipios (hasta el año 2010) es $n = 2456$, además el número de enfermos en el municipio s_i será denotado con y_i.

En este trabajo estamos interesados en modelar el riesgo de padecer obesidad a través del número de casos de obesidad registrados por municipio y algunas covariables, y entonces el vector $\boldsymbol{y} = (y_1, ..., y_n)$ denotará a los números de casos de obesidad en el municipio i.

El número esperado de casos de obesidad en el municipio i será denotado por c_i y éstos valores tuvieron que ser inferidos para todos los municipios a partir de los datos de la ENSANUT.

Las cantidades c_i son calculadas a partir de las tasas de obesidad por edad y sexo aplicadas a la población en riesgo dividida por edad y sexo. Sin embargo, estas cantidades c_i no están disponibles a nivel municipal para todo México debido a la falta de información en la base de datos por lo que tuvieron que ser estimadas a través de la prevalencia estatal de obesidad sin considerar la edad y el sexo, y ser ponderadas por las proporciones de los grupos según la edad y sexo (ver detalles más adelante).

Debido a que la obesidad no se considera una enfermedad contagiosa ni inusual entonces es posible asumir que el número de enfermos en cada municipio es mutuamente independiente y además siguen una distribución Poisson. Entonces es razonable pensar que:

$$y_i \sim Poi\left(c_i e^{x_i}\right) \text{ para } i = 1, ..., n$$

donde x_i representa el logaritmo del riesgo relativo en el municipio i y además, el modelo planteado para x_i es:

$$x_i = \boldsymbol{z_i}^t \boldsymbol{\beta} + u_i + v_i, \quad \text{para } i \in \{1, ..., n\} \tag{2.18}$$

donde

1. $x_i = \log(r_i)$ y r_i es el riesgo relativo en el i-ésimo municipio.
2. y_i denota el número de casos de obesidad en el municipio i.

3. $\boldsymbol{z_i}$ es el vector de covariables con $\boldsymbol{z_i} \in R^p$

4. $\boldsymbol{\beta}$ es el vector de coeficientes asociados a las covariables que incluyen al sexo y edad, con $\boldsymbol{\beta} \in \mathbb{R}^p$.

5. u_i denota el efecto de la estructura espacial sobre el logaritmo del riesgo relativo en la zona i.

6. v_i representa el efecto aleatorio por variables subyacentes o sin estructura alguna.

En el modelo (2.18), los parámetros sobre los cuales tenemos que hacer inferencia son $\boldsymbol{\beta}$, $\boldsymbol{u}$ y $\boldsymbol{v}$. Sea $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v})$ entonces por teorema de Bayes

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta})$$

donde $\boldsymbol{y} = (y_1, ..., y_n)$ y por el supuesto de independencia condicional se tiene que la verosimilitud es:

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|x_i) \propto \prod_{i=1}^{n} \exp(y_i x_i - c_i \exp(x_i)) = \exp(\boldsymbol{y}^t \boldsymbol{x} - \boldsymbol{c}^t \exp(\boldsymbol{x}))$$

donde $f(y_i|x_i)$ denota a la función de masa de probabilidad de una variable aleatoria Poisson con tasa $c_i \exp(x_i)$ y x_i el logaritmo del riesgo relativo en el municipio i que se modela conforme a (2.18).

Más aún, la distribución a priori toma en cuenta las dependencias y/o similitudes espaciales entre los municipios por lo que el modelar esas dependencias a través de campos aleatorios markovianos es razonable.

De acuerdo con Besag [13], una familia clásica de funciones que se han sido propuestas en el área de estadística espacial para la distribución del parámetro $\boldsymbol{u}$ son:

$$\pi(\boldsymbol{u}) \propto \exp\left(-\sum_{i<j} w_{ij}\phi(u_i - u_j)\right), \quad \boldsymbol{u} \in R^n \tag{2.19}$$

donde $\phi(z)$ es una función par y w_{ij} son pesos no negativos que denotan la asociación entre los municipios i y j. En esta aplicación pondremos $w_{ij} = 1$ si los municipios i y j son vecinos y hacemos $w_{ij} = 0$ en caso contrario (notar que $w_{ij} = w_{ji}$) y se elegirá a la función $\phi(z) = \frac{z^2}{2k}$ con k real positivo para la expresión (2.19).

Dicho esto, la distribución condicional de $\boldsymbol{u}$ condicionado al hiperparámetro k es:

$$\pi(\boldsymbol{u}|k) \propto \frac{1}{k^{(n-1)/2}} \exp\left(-\frac{1}{2k}\sum_{i\sim j}(u_i - u_j)^2\right) = \frac{1}{k^{(n-1)/2}} \exp\left(\boldsymbol{u}^t \boldsymbol{R}\boldsymbol{u}\right) \quad (2.20)$$

y de manera más precisa, $\boldsymbol{u}$ es un *campo aleatorio markovinano intrínseco de primer orden*[12], donde la notación $i \sim j$ significa que el municipio i y j son vecinos y $\boldsymbol{R}$ es la matriz dispersa de que captura estructura espacial [15] tal que:

$$R_{ij} = \begin{cases} n_i, & i = j \\ -1, & i \sim j \\ 0, & \text{otro caso} \end{cases}$$

donde n_i denota el número de vecinos del municipio i. De la expresión (2.20) se deriva la distribución condicional (2.17) del componente espacial u_i dado la matriz de adyacencias (binaria) el parámetro k, esto es:

$$u_i|\boldsymbol{u}_{-i}, k \sim N\left(\bar{u_i}, \frac{k}{n_i}\right) \quad (2.21)$$

donde $\bar{u_i}$ denota al promedio de los u_j tal que $j \sim i$ y n_i es el número de vecinos del vértice i como se había mencionado antes. En la expresión (2.21) se puede observar claramente que el efecto espacial u_i depende de los valores observados en vecinos de i.

Por otro lado, el vector $\boldsymbol{v}$ se asumirá como una realización de ruido gaussiano blanco con varianza $\lambda > 0$ desconocida por lo que la distribución a priori de $\boldsymbol{v}$ dado el hiperparámetro λ es:

$$\pi(\boldsymbol{v}|\lambda) \propto \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n} v_i^2\right) \quad (2.22)$$

Una vez establecidas las distribuciones para los parámetros podemos implementar el Gibbs sampling a través de las distribuciones completas de los parámetros.

2.4.3. Derivación de las Condicionales Completas:

Explícitamente, el modelo (2.18) es:

$$x_i = \sum_{j=1}^{p} z_{ij}\beta_j + u_i + v_i$$

donde p es el número de parámetros del modelo asociadas al intercepto y covariables (ya sean continuas o categóricas).

Como asumimos que y_i se distribuye como una variable Poisson con parámetro $c_i e^{x_i}$ y los parámetros del modelo (2.18) son

$$\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v})$$

con $\boldsymbol{\beta} \in \mathbb{R}^p$ y $\boldsymbol{u}, \boldsymbol{v}$ vectores en $\mathbb{R}^n$ donde n es el número de municipios que en nuestra aplicación es $n = 2556$ ya que es el número de municipios que poseen información para incluirla en el modelo.

La formulación del Teorema de Bayes dice que

$$\pi(\boldsymbol{\theta}|y) \propto L(\boldsymbol{\theta}; y)\pi(\boldsymbol{\theta})$$

que llega a ser:

$$Ł(\boldsymbol{\theta}; y) = \exp(\boldsymbol{y}^t \boldsymbol{x} - \boldsymbol{c}^t \exp(\boldsymbol{x}))$$

y retomando la distribuciones (2.20) y (2.22)

$$\pi(\boldsymbol{u}|k) \propto= k^{(n-1)/2} \exp\left(-\frac{k}{2}\boldsymbol{u}^t \boldsymbol{R}\boldsymbol{u}\right)$$

$$\pi(\boldsymbol{v}|\lambda) \propto \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\boldsymbol{v}^t \boldsymbol{v}\right)$$

y estableciendo distribuciones iniciales gama para los hiperparámetros de precisión k y λ asociados $\boldsymbol{u}$ y $\boldsymbol{v}$ respectivamente. Es decir si,

$$k \sim Gama(a_k, b_k)$$

$$\lambda \sim Gama(a_\lambda, b_\lambda)$$

entonces el modelo resultante es el clásico modelo de Besag[13]. La distribución inicial que se usará para los coeficientes asociados a las covariables es una normal multivariada con vector de medias cero y matriz de varianzas diagonal, i.e

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\beta) \tag{2.23}$$

Los valores de a_k, b_k, a_λ y b_λ así como los elementos de la matriz diagonal $\boldsymbol{\Sigma}_\beta$ serán fijos.

Entonces la distribución posterior es:

$$\begin{aligned}\pi\left(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{v},k,\lambda|\boldsymbol{y},\boldsymbol{a},\boldsymbol{b}\right) \propto & L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\beta}) \\ & \times\pi(\boldsymbol{u}|k)\pi(\boldsymbol{v}|\lambda)\pi(k|a_k,b_k)\pi(\lambda|a_\lambda,b_\lambda) \\ & = \exp\left[\boldsymbol{y}^t\boldsymbol{x}-\boldsymbol{c}^t\exp(\boldsymbol{x})\right]\exp\left[-\frac{1}{2}\boldsymbol{\beta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right] \\ & \times k^{(n-1)/2}\exp\left(-\frac{k}{2}\boldsymbol{u}^t\boldsymbol{R}\boldsymbol{u}\right)\lambda^{n/2}\exp\left[-\frac{\lambda}{2}\boldsymbol{v}^t\boldsymbol{v}\right] \\ & \times k^{a_k-1}\exp(-b_k k)\lambda^{a_\lambda-1}\exp(-b_\lambda\lambda)\end{aligned} \tag{2.24}$$

donde $\boldsymbol{a}=(a_k,b_k)$ y $\boldsymbol{b}=(a_\lambda,b_\lambda)$.

La distribución de la cual se pretende obtener una muestra es:

$$\begin{aligned}\pi\left(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{v},k,\lambda|\boldsymbol{y},\boldsymbol{a},\boldsymbol{b}\right) \propto & \exp\left[\boldsymbol{y}^t\boldsymbol{x}-\boldsymbol{c}^t\exp(\boldsymbol{x})-\frac{1}{2}\boldsymbol{\beta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}-\frac{k}{2}\boldsymbol{u}^t\boldsymbol{R}\boldsymbol{u}-\frac{\lambda}{2}\boldsymbol{v}^t\boldsymbol{v}-b_k k-b_\lambda\lambda\right] \\ & \times k^{a_k+\frac{(n-1)}{2}-1}\lambda^{a_\lambda+\frac{n}{2}-1}\end{aligned} \tag{2.25}$$

En la ecuación (2.25) se observa que las distribuciones a priori de los parámetros $\boldsymbol{\beta}$, $\boldsymbol{u}$ y $\boldsymbol{v}$ dado los parámetros k y λ conforman un campo aleatorio Markoviano. Sin embargo, $\boldsymbol{x}$ depende de la suma de tres cantidades i.e $\boldsymbol{x}=\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{u}+\boldsymbol{v}$ lo cual no es adecuado para la identificación de los parámetros.

Una manera de resolver esto acorde a Havard [12] es a través de reparametrizar la distribución 2.25 así que en lugar de generar una muestra de $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{v})$ se buscará obtener una muestra de $\boldsymbol{\theta^*}=(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{x})$.

Bajo esta nueva parametrización la distribución posterior es

$$\begin{aligned}\pi(\boldsymbol{\theta^*},k,\lambda|\boldsymbol{y},\boldsymbol{a},\boldsymbol{b}) \propto & \exp\left[\boldsymbol{y}^t\boldsymbol{x}-\boldsymbol{c}^t\exp(\boldsymbol{x})-\frac{1}{2}\boldsymbol{\theta^*}\boldsymbol{Q}\boldsymbol{\theta^{*t}}\right] \\ & \times\ k^{a_k+\frac{(n-1)}{2}-1}\lambda^{a_\lambda+\frac{n}{2}-1} \\ & \times\exp\left[-b_k k-b_\lambda\lambda\right]\end{aligned} \tag{2.26}$$

donde

$$Q=\begin{pmatrix}\boldsymbol{\Sigma_\beta}^{-1}+\lambda\boldsymbol{I} & \lambda\boldsymbol{1^t} & -\lambda\boldsymbol{1^t} \\ \lambda\boldsymbol{I} & k\boldsymbol{R}+\lambda\boldsymbol{I} & -\lambda\boldsymbol{I} \\ -\lambda\boldsymbol{I} & -\lambda\boldsymbol{I} & \lambda\boldsymbol{I}\end{pmatrix}$$

y $\boldsymbol{R}$ es la matriz de estructura del campo markoviano de primer orden con

$$R_{ij}=\begin{cases}n_i, & i=j \\ -1, & i\sim j \\ 0, & \text{otro caso}\end{cases}$$

donde n_i denota al número de vecinos del vértice i.

Derivado de la expresión (2.26) se tiene que las distribuciones posteriores de los parámetros k y λ son

$$\pi(k|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{x},\boldsymbol{y},\boldsymbol{a},\boldsymbol{b}) \propto k^{a_k+\frac{(n-1)}{2}-1}\exp\left(-b_k k-\frac{k}{2}\boldsymbol{u}^t\boldsymbol{R}\boldsymbol{u}\right) \tag{2.27}$$

$$\pi(\lambda|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{x},\boldsymbol{y},\boldsymbol{a},\boldsymbol{b}) \propto \lambda^{a_\lambda+\frac{n}{2}-1}\exp\left(-b_\lambda \lambda-\frac{\lambda}{2}\boldsymbol{v}^t\boldsymbol{v}\right) \tag{2.28}$$

que son los kernels de distribuciones gama, es decir

$$\begin{aligned}\pi(k|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{x},\boldsymbol{y},\boldsymbol{a},\boldsymbol{b}) &\sim Gama\left[a_k+\frac{n-1}{2}, b_k+\frac{1}{2}\boldsymbol{u}^t\boldsymbol{R}\boldsymbol{u}\right]\\ \pi(\lambda|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{x},\boldsymbol{y},\boldsymbol{a},\boldsymbol{b}) &\sim Gama\left[a_\lambda+\frac{n}{2}, b_\lambda+\frac{1}{2}\boldsymbol{v}^t\boldsymbol{v}\right]\end{aligned} \tag{2.29}$$

La estrategia para simular las muestras de la distribución posteriori 2.25 consiste en:

1. Generar una muestra de las precisiones k y λ en el primer con las distribuciones 2.29.

2. Generar muestras de $\boldsymbol{\theta^*}=(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{x})$ de manera conjunta.

Hay diferentes algoritmos para generar la muestra. Havard[12] propone el uso del "one block algorithm", en este trabajo optaremos por el camino de la aproximación normal para generar la muestra de las distribuciones.

Para generar muestras de $\boldsymbol{\theta^*}$ dado los demás parámetros se empleará el algoritmo de Metropolis-Hastings, por lo que se debe proponer una distribución candidata de

$$\pi(\boldsymbol{\theta^*}|\boldsymbol{y},k,\lambda)=\exp\left[\boldsymbol{y}^t\boldsymbol{x}-\boldsymbol{c}^t\exp(\boldsymbol{x})-\frac{1}{2}\boldsymbol{\theta^{*t}}\boldsymbol{Q}\boldsymbol{\theta^*}\right]$$

En la expresión anterior se omitieron los vectores fijos $\boldsymbol{a}$ y $\boldsymbol{b}$ para simplificar la notación.

La expansión de Taylor de segundo grado de $\boldsymbol{y}^t\boldsymbol{x}-\boldsymbol{c}^t\exp(\boldsymbol{x})$ alrededor de $\tilde{\boldsymbol{x}}$, es:

$$\boldsymbol{y}^t\boldsymbol{x}-\boldsymbol{c}^t\exp(\boldsymbol{x})\approx\boldsymbol{x}^t\boldsymbol{b}(\tilde{\boldsymbol{x}})-\frac{1}{2}\boldsymbol{x}^t\mathrm{diag}(\boldsymbol{s}(\tilde{\boldsymbol{x}}))\boldsymbol{x}$$

donde $\boldsymbol{s(\tilde{x})} = \boldsymbol{c}\exp(\tilde{\boldsymbol{x}})^t$ y $\boldsymbol{b(\tilde{x})} = \boldsymbol{y} + (\tilde{\boldsymbol{x}} - \boldsymbol{1})\boldsymbol{s(\tilde{x})}^t$. Entonces la densidad propuesta es

$$\begin{aligned} q(\boldsymbol{\theta^*}, \tilde{\boldsymbol{x}}|\boldsymbol{y}, k, \lambda) &= \exp\left[\boldsymbol{x}^t\boldsymbol{b(\tilde{x})} - \frac{1}{2}\boldsymbol{x}^t\text{diag}(\boldsymbol{s(\tilde{x})})\boldsymbol{x} - \frac{1}{2}\boldsymbol{\theta^{*t}Q\theta^*}\right] \\ &= \exp\left[\boldsymbol{x}^t\boldsymbol{b(\tilde{x})} - \frac{1}{2}\boldsymbol{\theta^{*t}V\theta^*}\right] \end{aligned}$$

con

$$V = \begin{pmatrix} \boldsymbol{\Sigma_\beta}^{-1} + \lambda\boldsymbol{I} & \lambda\boldsymbol{1^t} & -\lambda\boldsymbol{1^t} \\ \lambda\boldsymbol{I} & k\boldsymbol{R} + \lambda\boldsymbol{I} & -\lambda\boldsymbol{I} \\ -\lambda\boldsymbol{I} & -\lambda\boldsymbol{I} & \lambda\boldsymbol{I} + diag(\boldsymbol{s(\tilde{x})}) \end{pmatrix}$$

donde claramente $q(\boldsymbol{\theta^*}, \tilde{\boldsymbol{x}}) = N_C\left(\boldsymbol{b(\tilde{x})}, \boldsymbol{V}\right)$ en la parametrización canónica. Si $\boldsymbol{\theta^*} = (\boldsymbol{\beta^*}, \boldsymbol{u^*}, \boldsymbol{x^*})$ es el candidato, entonces lo aceptamos con probabilidad

$$\alpha = \text{mín}\left\{1, \frac{\pi(\boldsymbol{\beta^*}, \boldsymbol{u^*}, \boldsymbol{x^*}|\boldsymbol{y}, k, \lambda)}{\pi(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{x}|\boldsymbol{y}, k, \lambda)} \frac{q(\boldsymbol{\theta}, \tilde{\boldsymbol{x^*}}|k, \lambda)}{q(\boldsymbol{\theta^*}, \tilde{\boldsymbol{x}}|k, \lambda)}\right\}$$

2.5. Modelos lineales con variables categóricas

Esencialmente, un modelo lineal está determinado por un vector de observaciones (respuestas) $\boldsymbol{y}$ y una matriz de variables explicatorias (matriz de diseño) $\boldsymbol{X}$. Por ejemplo, un modelo lineal con dos variables explicatorias cuantitativas $\boldsymbol{x_1}$ y $\boldsymbol{x_2}$ queda planteado de esta manera:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, ..., n \tag{2.30}$$

donde β_0 denota al intercepto y β_1, β_2 son los coeficientes asociados al las variables $\boldsymbol{x_1}$ y $\boldsymbol{x_2}$, y ϵ denota al error aleatorio que típicamente se asume normal con media 0 y varianza constante. En este caso, la matriz de diseño queda definida así:

$$\boldsymbol{X} = [\boldsymbol{1} \quad \boldsymbol{x_1} \quad \boldsymbol{x_2}]$$

donde $\boldsymbol{1}$ es una columna de $1's$. Entonces el modelo (2.30) se expresa en forma matricial como

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

El caso donde se introducen variables de tipo categórico (factores) lleva al tradicional análisis de varianza para estudiar los cambios en la respuesta dependiendo de los niveles del factor. Este tipo de análisis requiere una codificación especial para el factor de tal manera que los resultados sean interpretables. Por ejemplo, en el caso donde se introduce un sólo factor a de k niveles como

única variable explicatoria el modelo para el análisis de varianza puede ser especificado como:

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij}, \quad j = 1, ..., k; \quad i = 1, ..., n_j$$

donde k es el número de niveles del factor, α_j denota al efecto del factor a en el nivel j y n_j denota al número de réplicas del experimento bajo el nivel j. Asumiendo que el numero de observaciones es n, la matriz de diseño de este modelo es

$$\boldsymbol{X} = [\mathbf{1} \quad \boldsymbol{X_a}]$$

donde $\boldsymbol{X_a}$ es una matriz binaria de incidencias de tamaño $n \times k$ donde se especifica la pertenencia al los niveles del tratamiento. Este modelo está sobre parametrizado debido a que el rango de $\boldsymbol{X}$ es k en lugar de $k+1$. La manera de romper esta redundancia de parámetros es mediante reparametrizar el modelo tal que la nueva matriz de diseño $\boldsymbol{X^*}$ tenga rango k. Esto se puede lograr mediante una matriz de contrastes $\boldsymbol{C_a}$ para el factor a de tamaño $k \times (k-1)$, y de esta manera, la nueva matriz de diseño es:

$$\boldsymbol{X^*} = [\mathbf{1} \quad \boldsymbol{X_a C_a}]$$

Una condición necesaria aunque no suficiente para que la matriz $\boldsymbol{X^*}$ tenga rango k es que la matriz cuadrada $[\mathbf{1} \quad \boldsymbol{C_a}]$ sea no singular.

La matriz de diseño $\boldsymbol{X^*}$ define especifica un modelo lineal pero los parámetros involucrados a veces no son fáciles de interpretar. A pesar de esto, la relación entre los parámetros del modelo redundante y los parámetros del modelo especificado por la matriz $\boldsymbol{X^*}$ es dada por

$$\boldsymbol{\alpha} = \boldsymbol{C_a \alpha^*} \tag{2.31}$$

donde $\boldsymbol{\alpha}$ denota al vector de parámetros del modelo redundante y $\boldsymbol{\alpha^*}$ a parámetros bajo la reparametrización. Más aún, si $\boldsymbol{c_a}$ es un vector tal que diferente de $\mathbf{0}$ tal que $\boldsymbol{c_a^t C_a} = \mathbf{0}$, entonces puede ser visto que usar los parámetros $\boldsymbol{\alpha^*}$ permite estimar los parámetros originales $\boldsymbol{\alpha}$ sujetos a la restricción $\boldsymbol{c_a^t \alpha} = \mathbf{0}$ lo cual usualmente es suficiente para hacerlos únicos[16].

Modelos con términos de interacción

En la apartado anterior se mostró como se puede incluir un variable explicatoria de tipo categórica en un modelo lineal y como se debe reparametrizar el modelo para evitar problemas computacionales. Cuando dos factores de tipo categórico son incluidos el modelo lineal el término de interacción entre estos dos factores puede ser tomado en cuenta también por el modelo. Por ejemplo,

un modelo lineal con dos factores a y b de r y s niveles respectivamente y que incluya el término de interacción tiene la siguiente forma:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \tag{2.32}$$

donde μ es el parámetro de la ordenada al origen, α y β son los parámetros asociados a los factores a y b respectivamente. El parámetro γ representa la interacción de los factores a y b.

El modelo anterior tiene la siguiente matriz de diseño:

$$\boldsymbol{X} = [\mathbf{1} \quad \boldsymbol{X_a} \quad \boldsymbol{X_b} \quad \boldsymbol{X_a} : \boldsymbol{X_b}]$$

donde de manera similar al modelo sin interacciones, su matriz de diseño para el modelo no redundante es

$$\boldsymbol{X^*} = [\mathbf{1} \quad \boldsymbol{X_a C_a} \quad \boldsymbol{X_b C_b} \quad (\boldsymbol{X_a C_a}) : (\boldsymbol{X_b C_b})] \tag{2.33}$$

más aún, el producto de las matrices que corresponden al término de la interacción se puede descomponer como:

$$(\boldsymbol{X_a C_a}) : (\boldsymbol{X_b C_b}) = (\boldsymbol{X_a X_b}) : (\boldsymbol{C_a} \otimes \boldsymbol{C_b}) \tag{2.34}$$

donde $\otimes$ denota al producto de Kronecker. La relación (2.34) implica que

$$\boldsymbol{\gamma} = (\boldsymbol{C_a} \otimes \boldsymbol{C_b})\boldsymbol{\gamma^*} \tag{2.35}$$

es decir, los parámetros de la interacción $\boldsymbol{\gamma}$ del modelo sobreparametrizado se obtienen mediante multiplicar la matriz resultante del producto de Kronecker de las matrices de contraste por el paramétro $\boldsymbol{\gamma^*}$.

2.5.1. Interpretación de los parámetros

Como se mencionó antes el uso de factores en un modelo lineal lleva consigo la reparametrización del modelo lineal mediante matrices que contrastan a los niveles de un factor. Por ejemplo, dado que la matriz de contrastes $\boldsymbol{C}$ es rango completo y si algún contraste ortogonal es empleado (como contrastes de Helmert o polinomial), entonces se cumple que $\boldsymbol{C}$ posee inversa única por la izquierda denotada por $\boldsymbol{C^+}$ tal que:

$$\boldsymbol{\alpha^*} = \boldsymbol{C^+}\boldsymbol{\alpha}, \text{ donde } \boldsymbol{C^+} = \left(\boldsymbol{C}\boldsymbol{C^t}\right)^{-1}\boldsymbol{C^t} \tag{2.36}$$

además, el patrón dentro de la matriz $\boldsymbol{C^+}$ provee una interpretación de los parámetros del modelo redundante a través de los parámetros del modelo

reparametrizado. Por ejemplo, para un factor de 4 niveles que se va a contrastar vía Helmert, la matriz de contrastes $\boldsymbol{C}$ es:

$$\boldsymbol{C} = \begin{bmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{bmatrix}$$

y entonces

$$\boldsymbol{C}^{+} = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ -\frac{1}{6} & -\frac{1}{6} & \frac{1}{3} & 0 \\ -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{4} \end{bmatrix}$$

y retomando la expresión (2.36) con $\boldsymbol{\alpha}^t = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$ se obtiene $\boldsymbol{\alpha^{*}}^t = (\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*)$ con $\alpha_1^* = \frac{\alpha_2 - \alpha_1}{2}$, $\alpha_2^* = \frac{1}{3}\left(\alpha_3 - \frac{\alpha_2+\alpha_1}{2}\right)$ y $\alpha_3^* = \frac{1}{4}\left(\alpha_4 - \frac{\alpha_3+\alpha_2+\alpha_1}{3}\right)$, es decir, los parámetros α_j^* denotan a la comparación de del parámetro α_{j+1} contra el promedio de los parámetros precedentes escalados por $j+1$ o equivalentemente, es la comparación de la media del grupo o clase $j+1$ contra el promedio de las clases anteriores.
La interpretación para cuando se emplea una matriz de contrastes vía polinomios ortogonales es que las entradas del vector de parámetros $\boldsymbol{\alpha^{*}}$ son los coeficientes de un polinomio ortogonal de grado $r-1$ (provisto de que el factor tenga r niveles). Por ejemplo, para un factor con 4 niveles la matriz de contrastes C para los niveles de los factores vía un polinomio ortogonal es:

$$\boldsymbol{C} = \begin{bmatrix} -0.6708204 & 0.5 & -0.2236068 \\ -0.2236068 & -0.5 & 0.6708204 \\ 0.2236068 & -0.5 & -0.6708204 \\ 0.6708204 & 0.5 & 0.2236068 \end{bmatrix}$$

y retomando la ecuación (2.31)) se ve que los parámetros α_1^*, α_2^* y α_3^* serán los coeficientes asociados a x, x^2 y x^3 es decir la respuesta y se trata de explicar como $\alpha_1^* x + \alpha_2^* x^2 + \alpha_3^* x^3$. Más aún, los valores de los coeficientes α^* ayudan a entender si existe un efecto lineal, cuadrático o cúbico dependiendo de los niveles del factor.

2.6. Intercambiabilidad y Modelos jerárquicos

El análisis de datos desde el punto de vista bayesiano se puede resumir en tres etapas[17]:

- Establecer un modelo *probabilístico completo* es decir, una distribución de probabilidad conjunta para los valores observados y subyacentes del problema.

- Obtener e interpretar la distribución posteriori de los parámetros dado los datos observados.
- Evaluar si es razonable el modelo planteado.

En el enfoque bayesiano, un concepto que permite un planteamiento y análisis de la información tomando en cuenta las consideraciones anteriores es el concepto de *intercambibilidad*. Sea $\{y_1, ..., y_n\}$ un conjunto de valores reales observados. En el caso en que los sub índices no aporten información relevante es decir, el orden en que fueron recolectados u observados las $y_i^{'}s$ entonces esas cantidades aleatorias son intercambiables.

Definición 10 *Decimos que una sucesión de variables aleatorias $Y_1, ..., Y_n$ son intercambiables la distribución conjunta de $Y_1, ..., Y_n$ es invariante ante cualquier permutación de los índices es decir*

$$p(Y_1, ..., Y_n) = p(Y_{\pi(1)}, ..., Y_{\pi(n)})$$

para cualquier permutación π.

Las cadenas de Markov son un ejemplo donde la condición de intercambiabilidad no se satisface.

A pesar de que la definición de intercambiabilidad en variables aleatorias es un concepto fácil de comprender, este tiene implicaciones fuertes en la teoría bayesiana. Más aún, si $\{y_1, ..., \}$ es una sucesión real de valores aleatorios, entonces existe una modelo paramétrico $p(y|\boldsymbol{\theta})$ (que es función de los $y_i^{'}$s) cuando $n \to \infty$ identificado por el parámetro $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ y densidad $p(\boldsymbol{\theta})$ tal que[18]:

$$p(y_1, ..., y_n) = \int_{\Theta} \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d(\boldsymbol{\theta})$$

Si asumimos intercambiabilidad sobre una sucesión de observaciones, entonces cualquier subconjunto de ellos es una muestra aleatoria de algún modelo $p(y_i|\boldsymbol{\theta})$ y además existe una distribución a priori $p(\boldsymbol{\theta})$ que describe la información inicial sobre el parámetro $\boldsymbol{\theta}$ que identifica al modelo. Dado que la existencia de una distribución a priori para el parámetro $\boldsymbol{\theta}$ está garantizada y que la información observada debe provenir de algún modelo identificado por el parámetro $\boldsymbol{\theta}$, se tiene que:

$$p(\boldsymbol{\theta}|y_1, ..., y_n) = \frac{\prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(y_1, ..., y_n)}$$

Además, si las observaciones son condicionalmente independientes (se asume implícitamente que la muestra aleatoria proviene de algun modelo) entonces

son necesariamente intercambiables.

En el caso donde exista información adicional para los valores observados y_i y suponiendo que los y_i no son intercambiables pero (y_i, x_i) sí lo son; aún es posible construir un modelo conjunto para (y_i, x_i) o un modelo condicional para $y_i|x_i$. De hecho, la manera usual de modelar intercambiabilidad en presencia de covariables x_i es a través de independencia condicional i.e:

$$p(\theta_1, ..., \theta_n|x_1, ..., x_n) = \int \prod_{i=1}^{n} p(\theta_j|\phi, x_j)p(\phi|\boldsymbol{x})d\phi$$

con $\boldsymbol{x}^t = (x_1, ..., x_n)$, y con el parámetro ϕ desconocido por lo que debemos especificar una distribución para éste (de ahí la frase modelo jerárquico ya que en la última expresión podemos ver como la ley de θ_i depende de ϕ).

Entonces la distribución priori conjunta para esos parámetros es:

$$p(\boldsymbol{\theta}, \phi) = p(\boldsymbol{\theta}|\phi)p(\phi)$$

y por lo tanto

$$p(\boldsymbol{\theta}, \phi|\boldsymbol{y}) \propto p(\boldsymbol{\theta}, \phi)p(\boldsymbol{y}|\boldsymbol{\theta}, \phi) = p(\boldsymbol{\theta}, \phi)p(\boldsymbol{y}|\boldsymbol{\theta})$$

ya que $p(\boldsymbol{y}|\boldsymbol{\theta}, \phi)$ sólo depende de $\boldsymbol{\theta}$.

De lo anterior se llega a que

$$p(\boldsymbol{\theta}, \phi|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\phi)p(\phi) \tag{2.37}$$

donde en la ecuación (2.37) es claro como la incertidumbre del parámetro $\boldsymbol{\theta}$ es modelada a través del hiperparámetro ϕ y también se observa la dependencia "jerárquica" entre las leyes de probabilidad de los parámetros.

2.7. Índice de Moran

Una medida ampliamente utilizada en estadística espacial para medir la autocorrelación espacial es la prueba de Moran. La autocorrelación espacial es más compleja que la autocorrelación usual de series de tiempo debido a que la autocorrelación espacial puede ser de dimensión 2 o 3.

La I de Moran se define como:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

donde y_i denota a la variable de interés en el $i-$ésimo municipio, $\overline{y}$ es promedio de la variable de interés y w_{ij} es el peso de la arista que conecta a los municipios i y j. En este trabajo se tomó $w_{ij} = 1$ cuando i y j son vecinos y $w_{ij} = 0$ en caso contrario. Además, dado que en la I de Moran se sustrae la media a cada valor y_i entonces I captura el patrón espacial que es causado por los pesos w_{ij}. El valor esperado de la I de Moran bajo la hipótesis nula de no auto correlación espacial es:

$$E(I) = \frac{-1}{n-1}$$

donde claramente $E(I) \to 0$ cuando $n \to \infty$. Los valores mayores a $E(I)$ denotan auto correlación espacial positiva.

El valor de la I de Moran es útil para saber si el modelo ajustado logró capturar el patrón espacial a través de calcularla sobre los residuales de cada municipio. Es decir, si y_i y $\hat{y}_i$ denotan al numero de casos observados y el valor ajustado (media Poisson estimada para cada grupo de edad y sexo) en el municipio i, entonces el residual

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

no debería reflejar el patrón espacial cuando el modelo propuesto tuvo un ajuste razonable.

2.8. Selección de modelo

Una pregunta natural es conocer si el modelo empleado es el más adecuado o la mejor opción entre un conjunto de modelos. Una manera de priorizar a un modelo sobre otro es través de medidas de ajuste y complejidad (en términos de números de parámetros) [19]. El criterio de información de devianza (*Deviance Information Criterion*) trata de resolver la pregunta acerca de qué modelos deben ser preferidos en términos de una medida de ajuste (usualmente la devianza) y la complejidad.

Supongamos que se tienen valores observados y que dependen de valores desconocidos ϕ los cuales representan a valores parámetros o hiperparámetros así como variables latentes en diferentes niveles del modelo. La distribución conjunta de (y, ϕ) es

$$\pi(y, \phi) \propto \pi(y|\theta)\pi(\theta|\psi)\pi(\psi)$$

con $\phi = (\theta, \psi)$ y donde se asume que y es condicionalmente de ψ dado θ y de hecho los valores de θ son los que influyen directamente en las medias de y, y ψ son los hiperparámetros de las distribuciones de θ. El problema de la

calidad del ajuste puede es medido a través de la distribución del logaritmo de la verosimilitud de los datos [20], [19]:

$$D(\theta) = -2\log \pi(y|\theta) + 2\log f(y). \tag{2.38}$$

donde $f(y)$ es una función que depende únicamente de los datos y que no afecta la comparación de modelos. De hecho, la expresión 2.38 es llamada *devianza bayesiana* [19]. Para el caso en que Y pertenece a la familia exponencial de un parámetro con $\mathbb{E}(Y) = \mu(\theta)$ entonces la devianza saturada se obtiene definiendo $f(y) = \pi(y|\mu(\theta) = y)$, en decir, una función de masa de probabilidad cuyo valor esperado es el observado en y.

La distribución posterior de $D(\theta)$ se obtiene a partir de la distribución posterior $\pi(\theta|y)$, donde $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$. De esta manera, el ajuste de un modelo está resumido en la esperanza posterior de la devianza [19]:

$$\bar{D} = \mathbb{E}_{\theta|y}(D).$$

Por otro lado, la complejidad del modelo se puede medir a través del número efectivo de parámetros p_D, que se define como:

$$\begin{aligned} p_D =& \mathbb{E}_{\theta|y}(D) - D(\mathbb{E}_{\theta|y}) \\ =& \bar{D} - D(\bar{\theta}) \end{aligned} \tag{2.39}$$

y el criterio de la información de devianza se define como

$$\begin{aligned} CID =& \bar{D} + p_D \\ =& D(\bar{\theta}) + 2p_D \end{aligned} \tag{2.40}$$

En términos prácticos, el CID es calculado durante la simulación del MCMC mediante monitorear θ y $D(\bar{\theta})$ y al final de la simulación tomar la media muestral de D menos la devianza evaluada en la media muestral de θ. Entonces,los modelos que se deben preferir son aquellos que reportan valores pequeños en el CID.

Capítulo 3

APLICACIÓN

3.1. Descripción de los datos

Los datos provienen en su mayoría de la Encuesta Nacional de Salud 2012. Esta encuesta contiene 384,929 registros de personas a los cuales se les medio el peso, estatura (talla), sexo y edad. Dado que la enfermedad a modelar es la obesidad, se calculó el IMC para cada uno de los registros y a través del contraste con los valores de referencia (cuadros 3.1 y 3.2) de la OMS se pudo diagnosticar si una persona padece obesidad.

Cuadro 3.1: Valores de referencia del IMC para el diagnóstico de obesidad en el grupo de mujeres.

Edad	Bajo Peso	Normal	Sobre Peso	Obesidad
10	<13.5	16.6	>19	>22.6
11	<13.9	17.2	>19.9	>22.6
12	<14.4	18.0	>20.8	>25.0
13	<14.9	18.8	>21.8	>26.2
14	<15.4	19.6	>22.7	>27.3
15	<15.9	20.2	>23.5	>28.2
16	<16.2	20.7	>24.1	>28.9
17	<16.4	21.0	>24.5	>29.3
18	<16.4	21.3	>24.8	>29.5
19	<16.5	21.4	>25	>29.7

El modelo que se presenta para el riesgo relativo de la obesidad tendrá como factores de entrada el sexo y grupos de edad. El cuadro 3.3 muestra la segmentación pos grupo de edad y sexo así como el número de casos de obesidad y sobrepeso observados en la base de datos. Debido a que las unidades muestrales de la ENSANUT están a nivel estatal y no municipal, la definición

Figura 3.1: Mapa de México. El color azul denota a los municipios con información válida de la ENSANUT 2012.

de los grupos de edad presentada en el cuadro 3.3 se hizo basada en la información censal del Instituto Nacional para el Federalismo y Desarrollo Municipal (INAFED)[21].

El número total de registros válidos sin considerar al grupo de personas entre 0 y 3 años de edad es de $74,080$. El mapa de Figura 3.1 muestra a los 742 municipios que contienen a los $74,080$ registros válidos de la ENSANUT 2012. Es importante resaltar un par de aspectos a considerar:

1. EL modelo espacial propuesto tiene el objetivo de imputar y/o estimar el riesgo en los municipios donde no hubo observaciones válidas

2. El mapa de la Figura 3.1 no muestra la segmentación por grupo de edad y sexo, es decir, los municipios de todos los datos válidos están denotados en azul en esa figura.

Cuadro 3.3: Número de casos de obesidad y sobrepeso encontrados por sexo y grupo de edad.

Grupo	Hombres	Mujeres
3-5	Sanos: 2483 Sobrepeso: 718 Obesidad: 353	Sanos: 2744 Sobrepeso: 602 Obesidad: 269
6-14	Sanos: 7927 Sobrepeso: 199 Obesidad: 2107	Sanos: 7961 Sobrepeso: 550 Obesidad: 1545
15-17	Sanos: 2326 Sobrepeso: 0 Obesidad: 2107	Sanos: 2304 Sobrepeso: 1 Obesidad: 308
18-24	Sanos: 2306 Sobrepeso: 565 Obesidad: 550	Sanos: 2665 Sobrepeso: 713 Obesidad: 758
25-29	Sanos: 569 Sobrepeso: 563 Obesidad: 396	Sanos: 860 Sobrepeso: 841 Obesidad: 664
30-59	Sanos: 2199 Sobrepeso: 4182 Obesidad: 2996	Sanos: 2634 Sobrepeso: 5185 Obesidad: 6014
> 60	Sanos: 1237 Sobrepeso: 1452 Obesidad: 740	Sanos: 1074 Sobrepeso: 1419 Obesidad: 1475

3.2. Descripción de covariables

Cada municipio de la base de datos tiene disponible las siguientes covariables:

1. Tasa de diabetes: Tasa de muerte por diabetes por cada 1000 mil habitantes, en el municipio del encuestado[1].
2. IDH: Índice de desarrollo humano del municipio
3. Rezago nutricional: Índice de rezago nutricional por municipio.
4. Tasa no acceso a salud: Porcentaje de personas que no son derechohabientes del sector salud por municipio.
5. Ingreso per cápita: Ingreso per cápita por municipio en miles de pesos.

A partir de estas covariables se consideró el coeficiente de correlación de Pearson que junto al análisis visual (ver Figura 3.2) nos ayuda identificar ciertas asociaciones lineales entre las covariables contenidas en la ENSANUT, por ejemplo:

[1]Esta variable fue incluida para recuperar el patrón espacial que pudo ser afectado por el muestreo.

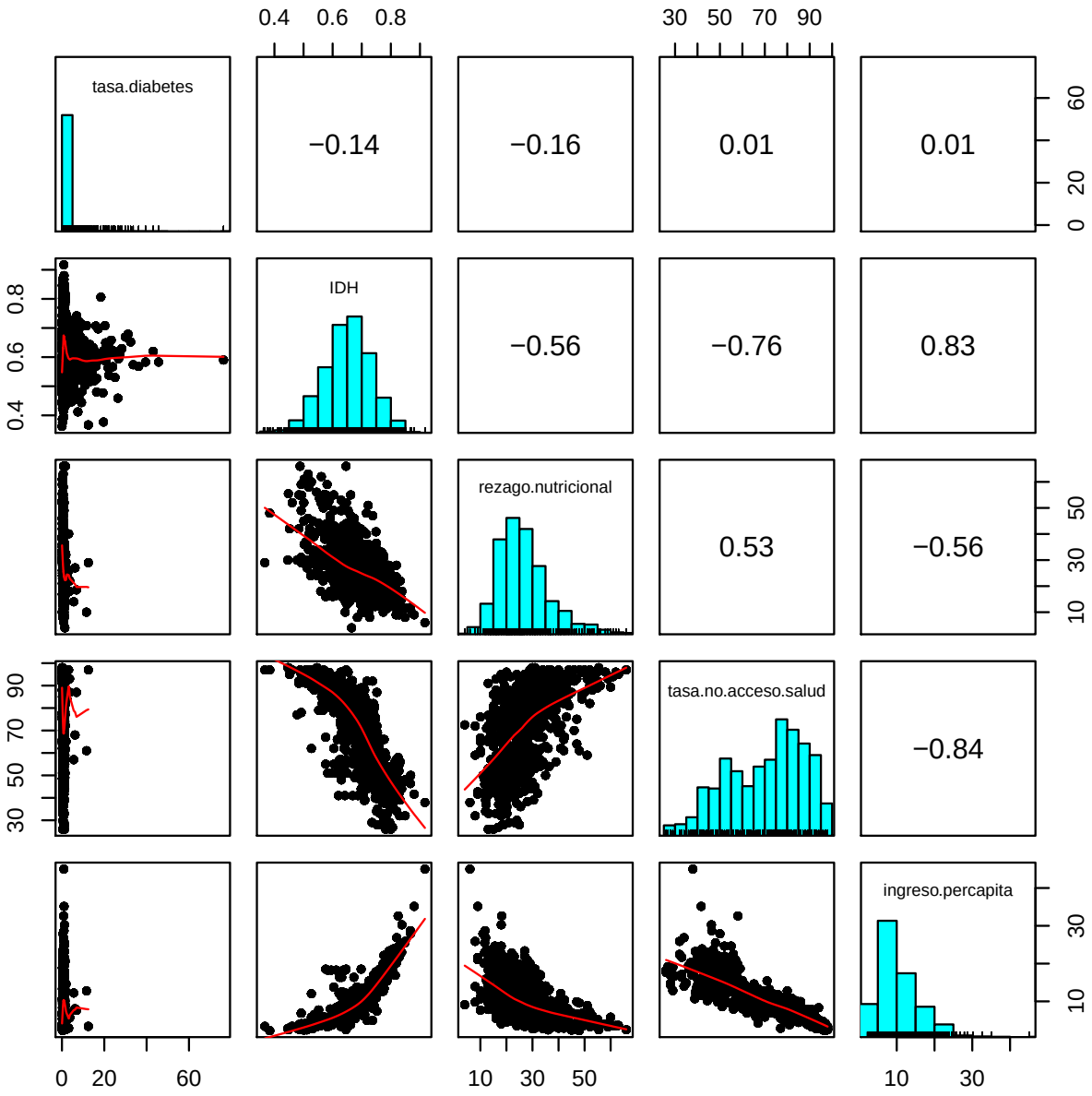


Figura 3.2: Gráfica de dispersión de las covariables socioeconómicas a nivel municipal reportadas en la ENSANUT 2012.

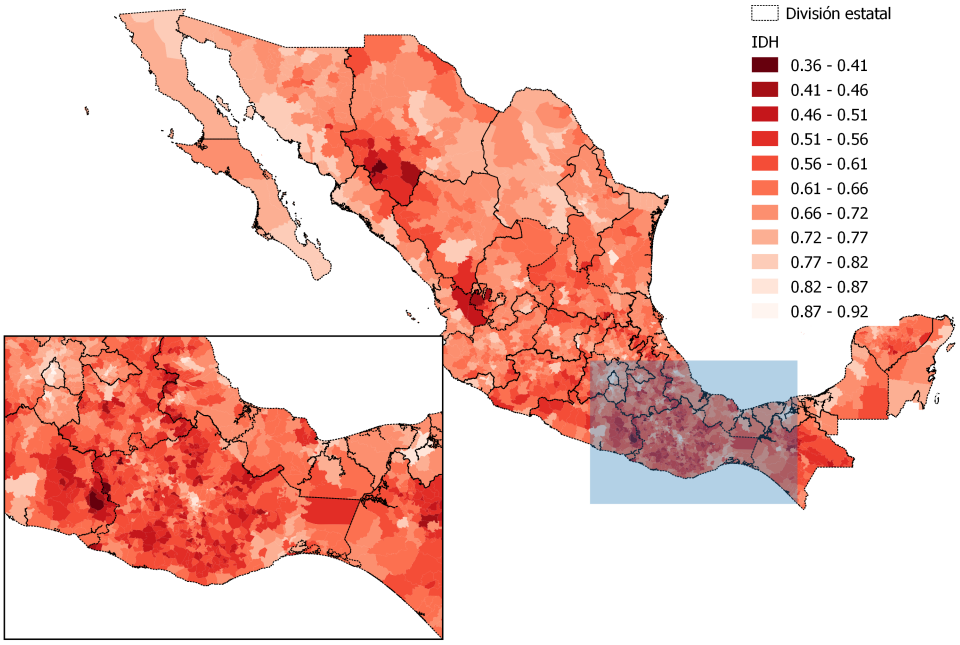


Figura 3.3: Mapa del índice de desarrollo humano (IDH) de México por municipio en 2010. Fuente: INAFED[21].

1. El IDH está correlacionado con el rezago nutricional, tasa no acceso a salud e ingreso per cápita.

2. La tasa de diabetes no muestra muestra asociación lineal con el resto de las covariables.

En ese sentido, las únicas covariables que se incluyeron como variables de entrada en el modelo fueron la tasa de diabetes por municipio (para intentar recuperar el patrón espacial que pudo ser afectado por el muestreo [2]) y el índice de desarrollo humano (IDH). Este último engloba varios aspectos socioeconómicos de lo municipios, y por esta razón se descartó a las variables rezago nutricional, tasa de no acceso a salud y el ingreso per cápita.

[2]Recordar que el muestreo fue diseñado de tal manera que un individuo encuestado representa a cierto número de personas a nivel estatal y no a nivel municipio donde vive el encuestado

3.2.1. Especificación del modelo

El modelo es Poisson es

$$Y_i \sim Poi(c_i x_i)$$

donde

1. Y_i es el número de casos de obesidad por grupo de edad y sexo en el municipio i.
2. c_i representa el número esperado de enfermos por grupo de edad y sexo en el municipio i.
3. x_i es el riesgo relativo en el municipio i.
4. n_i es la población del municipio i (por grupo de edad y sexo).

Como se mencionó antes, el diseño de muestra de la ENSANUT 2012 fue hecho a nivel estatal y no municipal, entonces los valores c_i (que están a nivel municipal) se calcularon de esta manera:

$$c_i = n_i \cdot p_{Est} \cdot p_{Gpo} \cdot p_i$$

donde p_{Est} es la prevalencia estatal de padecer obesidad sin importar grupo de edad o sexo, p_{Gpo} es la proporción de personas que pertenecen a cierto grupo de edad y sexo y p_i es la proporción de personas del estado que viven en el municipio i sin importar edad y sexo.

Los valores observados Y_i fueron obtenidos de las encuestas asumiendo que cada persona encuestada representa a un número de personas a nivel estatal mediante un ponderador reportado en la ENSANUT.

El número de personas observadas por grupo de edad y sexo con obesidad en el municipio i se define como:

$$Y_i = n_{pond_i} \cdot p_i$$

donde n_{pond_i} es la suma del número de personas (a nivel estatal) que son representadas por un encuestado que padece obesidad y que pertenece a cierto grupo de edad-sexo y que vive en el municipio i.

El valor p_i es la proporción de personas del estado que viven en el municipio i sin importar edad y sexo como ya se había mencionado antes.

Cuadro 3.4: Ejemplo del uso del ponderador. $n_{pond_a} = 100$ **y** $n_{pond_b} = 305$ **acorde a la expresión 3.1**

Encuestado	Obesidad	Ponderador	Municipio	Estado
1	1	100	a	Estado X
2	1	185	b	Estado X
3	0	91	a	Estado X
4	1	120	b	Estado X

El cuadro 3.4 ejemplifica el uso del ponderador. Para este ejemplo vamos a suponer que el Estado X tiene dos municipios (a y b), y que sólo hubo 4 personas encuestadas.

El encuestado número 1 (que pertenece a cierto grupo de edad y sexo) tiene obesidad (codificado con el valor 1 en la columna obesidad) y vive en el municipio a.

También hay 2 personas encuestadas (asumamos que pertenecen al mismo grupo de edad y sexo) que tienen obesidad y que viven en el municipio b. Entonces n_{pondef} toma los valores 100 y $185 + 120 = 305$ respectivamente, lo cual dice que los encuestados que padecen obesidad de los municipios a y b representan a 100 y 305 personas con obesidad a nivel estatal y que además pertenecen a cierto grupo de edad y sexo (dependiendo de la edad y sexo de los encuestados).

De manera general, el número de personas con obesidad y que pertenecen a un grupo específico de edad y sexo es:

$$n_{pond_i} = \Sigma_{j \in \text{Municipio i}} \quad O_j \cdot pondef_j \tag{3.1}$$

donde $pondef_j$ es el número de personas a nivel estatal que representa el encuestado j (que vive en el municipio i y pertenece a un grupo de edad y sexo). $O_j \in \{0, 1\}$ es el estatus del encuestado (1 significa que el encuestado j padece obesidad, 0 el caso contrario).

Entonces el modelo propuesto para el logaritmo del riesgo relativo x_i del municipio i es:

$$\log x_i = \mu + \mu_1 \text{IDH}_i + \mu_2 \text{Diabetes}_i + Edad + Sexo + Edad \cdot Sexo + u_i + v_i \tag{3.2}$$

donde IDH_i y Diabetes_i son el índice de desarrollo humano y la tasa de muertes asociadas a la diabetes por cada 1000 habitantes en el municipio i respectivamente, mientras que $Edad$ y $Sexo$ son variables de entrada categóricas que

indican la pertenencia a cierto grupo de edad o sexo.

En el caso del factor *Edad* los niveles[3] que toma son $3-5$, $6-14$, $15-17$, $18-24$, $25-29$, $30-59$ y 60 años o más mientras que el factor *Sexo* sólo toma dos valores categóricos (hombre/mujer). Los contrastes ortogonales que se emplearán son Helmert y polinomial para el *grupo de edad* y *sexo* respectivamente.

Siguiendo la notación acorde a la expresión (2.32), las matrices de contraste para estos factores son:

$$\boldsymbol{C_a} = \begin{bmatrix} -0.7071068 \\ 0.7071068 \end{bmatrix} \tag{3.3}$$

$$\boldsymbol{C_b} = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 2 & -1 & -1 & -1 & -1 \\ 0 & 0 & 3 & -1 & -1 & -1 \\ 0 & 0 & 0 & 4 & -1 & -1 \\ 0 & 0 & 0 & 0 & 5 & -1 \\ 0 & 0 & 0 & 0 & 0 & 6 \end{bmatrix} \tag{3.4}$$

En el modelo (3.2) el término u_i denota el efecto de la estructura espacial mientras que v_i al efecto aleatorio y la suma $\psi_i = u_i + v_i$ es identificable.

En la gráfica 3.4 se reporta el número de casos de obesidad observados en la muestra de la ENSANUT 2012. Esta gráfica muestra de manera cualitativa que existe interacción entre los factores de *grupo de edad* y *sexo*. El contraste más notable entre estos factores se da en el grupo de edad de 60 años o más, ya que en el caso de los hombres hay muchos casos registrados de obesidad en comparación con las mujeres.

[3]los grupos de edad fueron definidos de esta manera para encajar con los que INEGI hace la segmentación. Personas menores de 3 años no fueron tomadas en cuenta para este estudio.

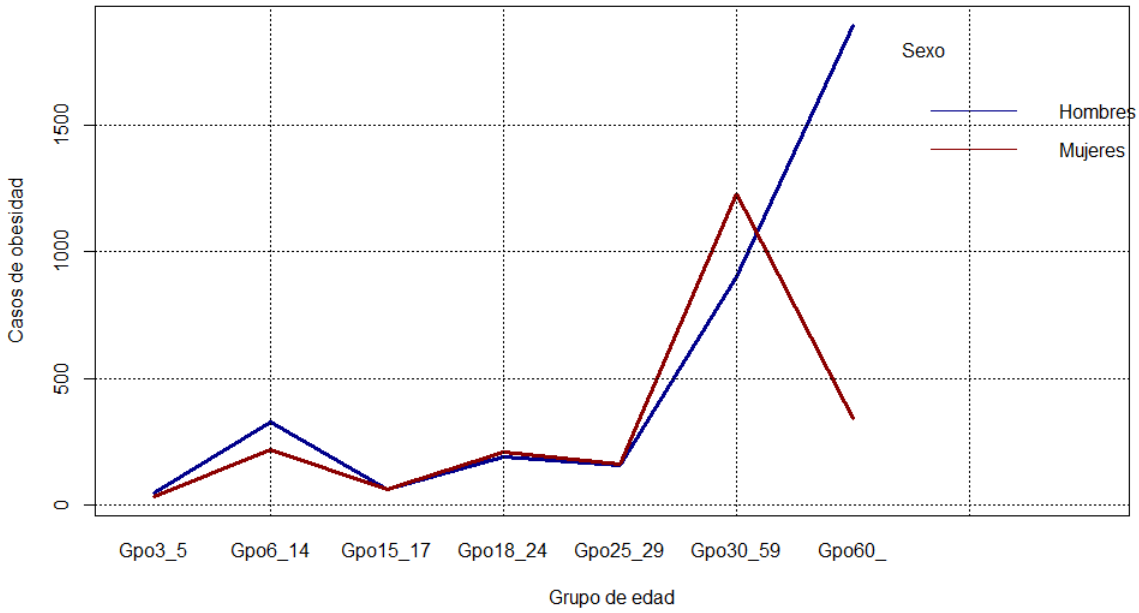


Figura 3.4: Gráfica de interacción del número de casos por municipio de obesidad de hombres y mujeres encontrados en la muestra de la ENSANUT 2012.

Capítulo 4

DISCUSIÓN Y RESULTADOS

4.1. Método

Los datos relacionados a áreas geográficas (que no se traslapan) están presentes en varios campos de estudio como la agricultura, educación, epidemiología, análisis de imágenes entre otros. En los casos mencionados anteriormente, lo típico es observar autocorrelación espacial es decir, las leyes de probabilidad de un área en particular son parecidas a las leyes de probabilidad de las áreas vecinas. La teoría dice que una proporción de la autocorrelación espacial puede ser explicada por covariables asociadas a cada área geográfica, sin embargo es posible que los residuales del modelo ajustado aún contengan parte de la estructura espacial. Lo típico a la hora de tratar con residuales autocorrelacionados es aumentar los predictores con un conjunto de efectos aleatorios autocorrelacionados a través de incluirlos en un modelo bayesiano jerárquico [14].Estos efectos aleatorios usualmente se representan en forma de un campo aleatorio markoviano con distribuciones a priori condicionales autoregresivas las cuales capturan la estructura espacial a través de la matriz de adjacencias de las áreas geográficas. En este sentido, Besag fue pionero en proponer estos modelos autorregresivos [13]. Sin embargo en los últimos 20 años han surgido otras propuestas como la de Leroux donde modelo de Besag-York-Mollie es un caso particular de esta formulación). El modelo de Leroux incluye un parámetro de dependencia espacial ρ que sigue una distribución uniforme entre 0 y 1 [22].

Un inconveniente al usar distribuciones condicionales autorregresivas es que éstas modelan la autocorrelación espacial de manera global y no local. Por ejemplo, en aplicaciones con datos de tipo económico puede haber pequeños conglomerados de zonas con un ingreso económico alto que están localizadas dentro de zonas con menor ingreso.

Existen propuestas para tratar con este tipo de discontinuidades y "suavizar" el efecto espacial de manera local y no global. Tales propuestas pueden fueron desarrolladas por Lee y Mitchell(2012)[23] y Lee y Serran(2015) [24].

En este trabajo se optó por la clásica formulación que fue propuesta por Besag-York-Mollie [13], es decir, el modelo utilizado aquí "suaviza" el efecto espacial de manera global y no local. Consecuentemente el objetivo es proveer un modelo estadístico que impute los riesgos relativos para los municipios que tienen observaciones perdidas.

4.2. Resultados (modelo y mapas)

A través del método de simulación Monte Carlo vía Cadenas de Markov, se generaron 15 millones de muestras y se guardaron las muestras cada 10 mil pasos para evitar autocorrelación con un periodo de calentamiento de un millón. Entonces los parámetros estimados con esta muestra y sus respectivos intervalos de probabilidad para el modelo (3.2) ya parametrizado con las matrices de contraste (3.3) y (3.4) se muestran en el Cuadro 4.1. En este cuadro

Cuadro 4.1: Resumen estadístico de las muestras generadas para los parámetros del modelo (3.2) vía MCMC de una corrida de 15 millones de iteraciones guardando cada 10 mil pasos con un periodo de calentamiento de 1 millón.

Parámetro	Covariable	$Mediana$	$Q_{0.025}$	$Q_{0.975}$	exp($Mediana$)
μ	Intercepto	1.070	0.972	1.125	2.915
μ_1	IDH	-5.945	-6.032	-5.732	0.002
μ_2	Diab	0.257	0.256	0.258	1.293
β_1^*	Gpo1	0.179	0.174	0.180	1.196
β_2^*	Gpo2	-0.012	-0.012	-0.010	0.988
β_3^*	Gpo3	0.061	0.060	0.062	1.063
β_4^*	Gpo4	0.0795	0.0795	0.0798	1.082
β_5^*	Gpo5	0.136	0.135	0.137	1.144
β_6^*	Gpo6	0.259	0.258	0.261	1.296
α_1^*	Sexo(Hombre)	-0.209	-0.211	-0.205	0.810
γ_1^*	Sexo(Hombre)*Gpo1	0.084	0.064	0.106	1.088
γ_2^*	Sexo(Hombre)*Gpo2	0.087	0.079	0.095	1.091
γ_3^*	Sexo(Hombre)*Gpo3	-0.047	-0.050	-0.045	0.953
γ_4^*	Sexo(Hombre)*Gpo4	-0.012	-0.014	-0.011	0.987
γ_5^*	Sexo(Hombre)*Gpo5	0.022	0.019	0.024	1.022
γ_6^*	Sexo(Hombre)*Gpo6	-0.163	-0.177	-0.148	0.848

el parámetro μ corresponde a la ordenada al origen (intercepto) en el modelo lineal, μ_1 representa al efecto del índice de desarrollo humano (IDH), μ_2 es el coeficiente asociado a la tasa de muertes por diabetes, β_i^* es el contraste de Helmert para los grupos de edad para $i = 1, ..., 6$ y α_1^* es el efecto del género donde la categoría basal es ser hombre.

El efecto de la interacción entre el factor *sexo* y *edad* es cuantificado por γ_i^* con $i = 1, ..., 6$, donde Gpo1 denota al contraste del grupo de edad 6-14 vs 3-5 años, Gpo2 al contraste 15-17 años vs grupos de edad anteriores y sucesivamente Gpo3 contrasta 18-24 años vs grupos de edad anteriores, Gpo4 a 25-29 años vs edades anteriores, Gpo5 a 30-59 años vs edades anteriores y Gpo6 a 60 o más años vs edades anteriores.

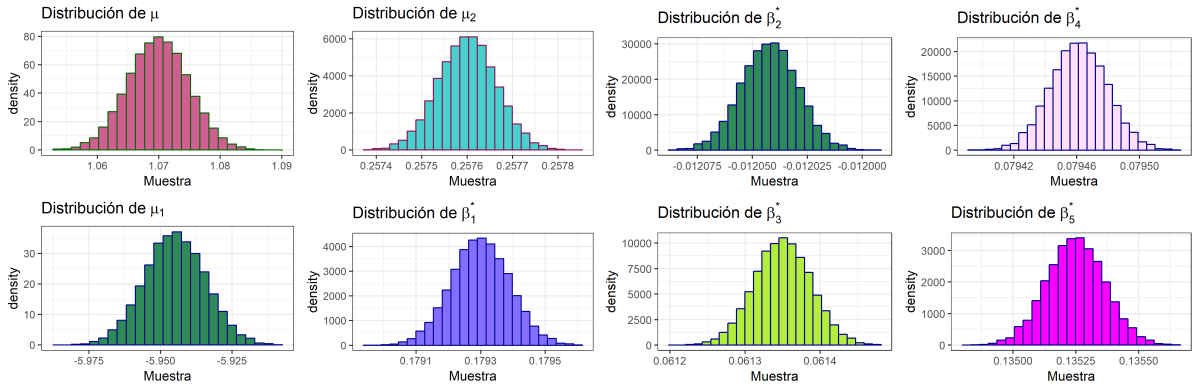
Como se puede observar en el cuadro 4.1, los intervalos definidos por los cuantiles de 0.025 y 0.975 de las distribuciones marginales posteriores no contienen al valor 0, por lo que podríamos concluir que hay un efecto de las covariables (y sus niveles) sobre el logaritmo del riesgo relativo. En el caso particular del parámetro μ_1 que está asociado al IDH, se ve que éste tiene una influencia negativa o inversa respecto al riesgo relativo, es de decir, a mayor IDH menor riesgo de padecer obesidad.

Este resultado es similar a lo reportado en otros estudios, por ejemplo, el estudio hecho por McLaren [25] menciona que para mujeres en países con alto IDH, la obesidad está asociada a personas con un estatus socioeconómico bajo y también en este mismo estudio se reportó que en países con IDH bajo, se mantuvo la asociación entre obesidad y estatus socioeconómico bajo mientras que en países con IDH medio hay resultados mixtos. El punto discrepancia con este trabajo del estudio hecho por McLaren es que su enfoque es a través de hacer una segregación de países por IDH para después considerar al estatus socioeconómico como un factor explicativo de la prevalencia de obesidad implicó utilizar el nivel educativo como medida de aproximación al este factor [26] mientras el enfoque presentado aquí toma directamente IDH para alimentar al modelo por que es una medida estándar que tiene la ventaja de incluir aspectos como la esperanza de vida, el producto interno bruto y los años de escolaridad. En el caso de los hombres, McLaren reporta que no hay asociación clara entre el estatus económico y la prevalencia obesidad para países con IDH alto y medio, mientras que en países con IDH bajo se reportó asociación positiva entre estatus socioeconómico y prevalencia de obesidad. Nuevamente, el factor estatus socioeconómico no explicó la prevalencia de obesidad en países de IDH alto y medio. La conclusión es que el IDH puede ser más robusto para explicar el riesgo de padecer obesidad en países menos desarrollados.

Otro estudio del año 2016 señaló que hay mayor prevalencia de obesidad para hombres y mujeres mayores de 18 años que viven en Estados Unidos que países en el sur de Asia. Este estudio también reporta que países del sur de Asia tienen una correlación positiva y significativa entre la prevalencia de padecer obesidad y el IDH (0.48 y 0.36 de correlación respectivamente) [27] lo cual es contrario a los resultados que obtuvimos pues nuestra interpretación es que a mayor IDH hay menor riesgo de padecer obesidad.

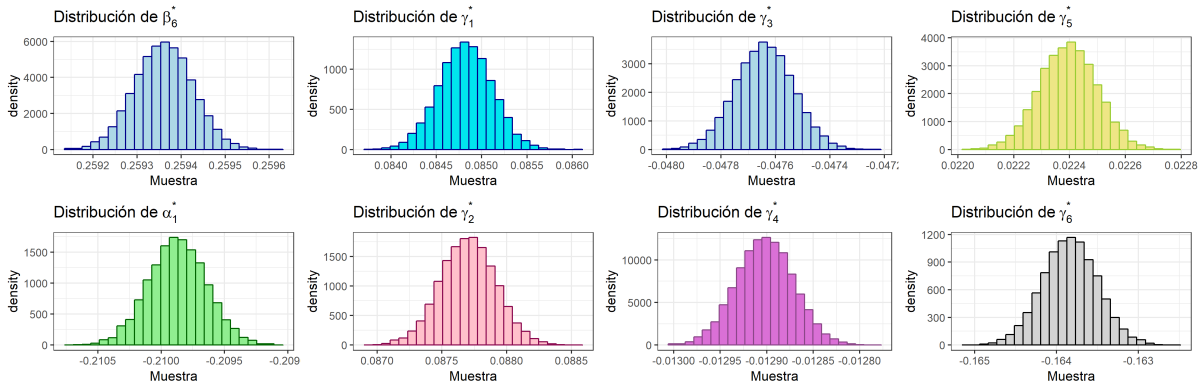
En el caso de América Latina no hay mucha información o estudios acerca del IDH y la prevalencia para contrastar los resultados obtenidos en este trabajo.

Las gráficas de las distribuciones posteriores marginales de las muestras obtenidas por simulación Monte Carlo vía Cadenas de Markov son reportadas en la Figura 4.1. Las trayectorias de la muestras generadas para los parámetros descritos en el cuadro 4.1 son reportadas en la Figura 4.2.



(a) Densidades de μ (intercepto), μ_1 (IDH), μ_2 (tasa diabetes) y β_1^* (contraste Gpo1)

(b) Densidades de β_2^* (Gpo2), β_3^* (Gpo3), β_4^* (Gpo4) y β_5^* (Gpo5).



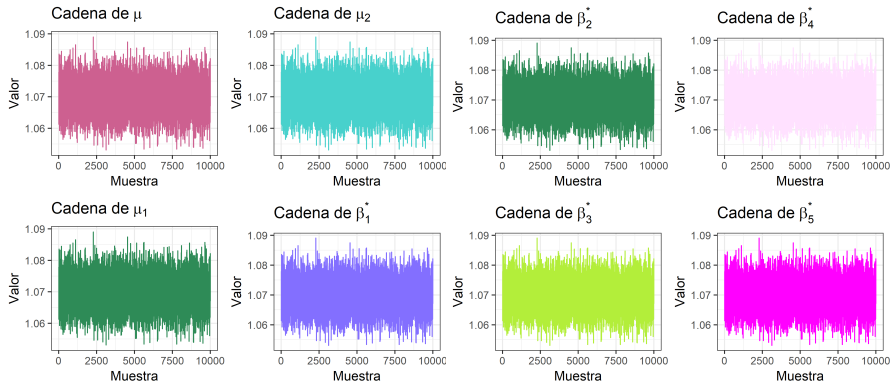
(c) Densidades de β_6^* (Gpo6), α_1^* (sexo), γ_1^* y γ_2^* (interacción Gpo1 y Gpo2 con sexo respectivamente).

(d) Densidades de γ_3^*, γ_4^*, γ_5^* y γ_6^* (interacciones Gpo3, Gpo4, Gpo5 y Gpo6 con sexo respectivamente)

Figura 4.1: Distribuciones posteriores marginales obtenidas con la muestra generada vía MCMC para los parámetros del Cuadro 4.1.

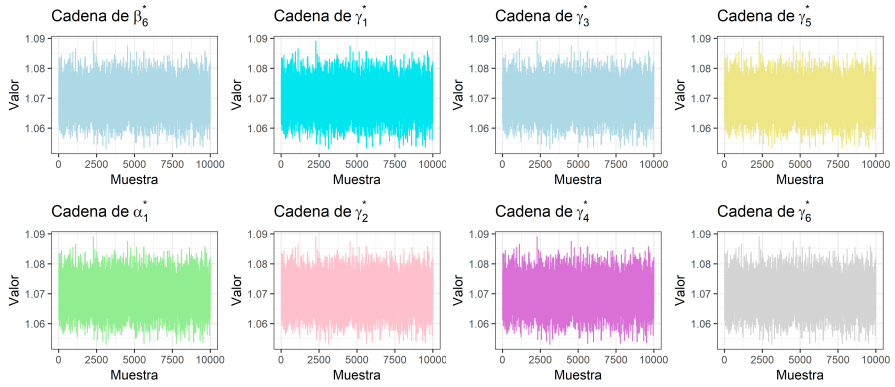
4.2.1. Interpretación y relación entre los parámetros estimados del modelo parametrizado y redundante

Las estimaciones de los parámetros reportados en el cuadro 4.1 muestran que existe un efecto de interacción entre el factor *sexo* y *grupo de edad* del modelo reparametrizado. En este sentido, para explicar la interacción de estos factores es conveniente revertir la parametrización para obtener los estimadores de los parámetros del modelo redundante y poder dar una interpretación un poco más clara de los parámetros estimados a través analizar la relación algebraica que existe entre el modelo redundante y el reparametrizado.



(a) Cadena de μ, μ_1,μ_2 y β_1^*

(b) Cadena de β_2^*, β_3^*, β_4^* y β_5^*.



(c) Cadena de β_6^*, α_1^*, γ_1^* y γ_2^*.

(d) Cadena de γ_3^*, γ_4^*, γ_5^* y γ_6^*

Figura 4.2: Trayectorias de las muestras generadas para los parámetros descritos en el Cuadro 4.1 obtenidas vía MCMC.

Entonces, si quisiéramos hacer comparaciones entre grupos de edad y sexo vía el modelo redundante basta con retomar la expresión (2.35) para así obtener los parámetros $\boldsymbol{\gamma}$ que corresponden a los términos de interacción entre el factor *sexo* y *grupo de edad*.

Estos se calculan de la siguiente manera:

$$\boldsymbol{\gamma} = \boldsymbol{C_a}\boldsymbol{\gamma^*}\boldsymbol{C_b^t}$$

donde $\boldsymbol{\gamma^*}$ es el vector de parámetros estimados que corresponden a las interacciones del modelo después de ser reparametrizado con las matrices de contraste (3.3) y (3.4).

Tomando $\boldsymbol{\gamma^*} = (\gamma_1^*, \gamma_2^*, \gamma_3^*, \gamma_4^*, \gamma_5^*, \gamma_6^*)$ (es decir, los valores reportados en el

cuadro (4.1)) y reemplazando en la expresión (4.2.1) se llega a que los parámetros del modelo redundante son:

Cuadro 4.2: Parámetros correspondientes a las interacciones en el modelo redundante

Sexo	Gpo3a5	Gpo6a14	Gpo15a17	Gpo18a24	Gpo25a29	Gpo30a59	Gpo60+
Hombres	γ_1	γ_3	γ_5	γ_7	γ_9	γ_{11}	γ_{13}
Mujeres	γ_2	γ_4	γ_6	γ_8	γ_{10}	γ_{12}	γ_{14}

Cuadro 4.3: Parámetros estimados correspondientes a las interacciones del modelo redundante (2.32)

Sexo	Gpo3a5	Gpo6a14	Gpo15a17	Gpo18a24	Gpo25a29	Gpo30a59	Gpo60+
Hombres	-0.0207	-0.1407	-0.2667	-0.0081	-0.0634	-0.1950	0.6949
Mujeres	0.0207	0.1407	0.2667	0.0081	0.0634	0.1950	-0.6949

En este caso, se puede corroborar la condición de que la suma de los parámetros estimados (por filas y columnas) es 0. Un aspecto importante es que se puede interpretar las interacción entre *grupo de edad* y *sexo* ya sea con los γ_i's o los γ_i^*'s. Aquí se optará por la segunda opción.

Interpretación de los coeficientes del modelo reparametrizado

El cuadro (4.1) mostró que existe una interacción entre los factores *grupo de edad* y *sexo*. Entonces, seguido del *principio de marginalidad* haremos caso omiso de los efectos principales, es decir, omitiremos la interpretación de los coeficientes β_i^*'s y α_1^*, por lo que nos enfocaremos en interpretar directamente los coeficientes γ_i^*'s.

Para hacer dicha interpretación, es conveniente escribir de manera explícita la relación entre los coeficientes $\boldsymbol{\gamma}$ y $\boldsymbol{\gamma^*}$. Recordando que la relación algebraica que hay entre los parámetros de interacción del modelo redundante y el modelo reparametrizado en presencia de dos factores categóricos a y b es:

$$\boldsymbol{\gamma} = (\boldsymbol{C_b} \otimes \boldsymbol{C_a})\boldsymbol{\gamma^*} \tag{4.1}$$

donde $\boldsymbol{C_a}$ y $\boldsymbol{C_b}$ son las matrices de contraste de a y b y acordé a lo expuesto en la expresión (2.36) se llega a que

$$\boldsymbol{\gamma^*} = (\boldsymbol{C_b} \otimes \boldsymbol{C_a})^+\boldsymbol{\gamma} \tag{4.2}$$

entonces, los parámetros $\boldsymbol{\gamma}$ que corresponden a las interacciones en el modelo redundante multiplicado por la inversa generalizada del producto de Kronecker de las matrices de contraste dan como resultado a los parámetros estimados de modelo ya reparametrizado que son los que se mostraron en el cuadro (4.1).

De manera explícita para el caso en que $\boldsymbol{C_a}$ y $\boldsymbol{C_b}$ son los contrastes de polinomios ortogonales y Helmert, la relación algebraica que se hay entre $\boldsymbol{\gamma}$ y $\boldsymbol{\gamma^*}$ es:

$$\begin{aligned}
\gamma_1^* &= \frac{2\cos(\pi/4)}{2}\left[\gamma_4 - \gamma_2\right], \\
\gamma_2^* &= \frac{2\cos(\pi/4)}{3}\left[\gamma_6 - \frac{\gamma_2+\gamma_4}{2}\right], \\
\gamma_3^* &= \frac{2\cos(\pi/4)}{4}\left[\gamma_8 - \frac{\gamma_2+\gamma_4+\gamma_6}{3}\right], \\
\gamma_4^* &= \frac{2\cos(\pi/4)}{5}\left[\gamma_{10} - \frac{\gamma_2+\gamma_4+\gamma_6+\gamma_8}{4}\right], \\
\gamma_5^* &= \frac{2\cos(\pi/4)}{6}\left[\gamma_{12} - \frac{\gamma_2+\gamma_4+\gamma_6+\gamma_8+\gamma_{10}}{5}\right], \\
\gamma_6^* &= \frac{2\cos(\pi/4)}{7}\left[\gamma_{14} - \frac{\gamma_2+\gamma_4+\gamma_6+\gamma_8+\gamma_{10}+\gamma_{12}}{6}\right],
\end{aligned} \tag{4.3}$$

Las relaciones mostradas en las ecuaciones (4.3) dicen como se pueden obtener los parámetros γ_i^* en términos de los parámetros γ_j (que se reportaron en el cuadro 4.2). Es claro que las interacciones γ_i^* representan el contraste de cierto grupo de edad versus el promedio de los grupos de edad anteriores cuando el *sexo* es mujer[1].

Consecuentemente, γ_1^* representa el efecto sobre el riesgo relativo al pasar del grupo de edad de 3 a 5 años al de 6 a 14 años siendo mujer (recordar que el nivel basal del factor sexo es *hombre*), lo cual significa que existe un aumento del 8.8 % en el riesgo relativo (ver valor exponenciado de γ_1^* en el cuadro 4.1).

El parámetro γ_2^* sugiere que el riesgo relativo aumenta un 9.1 % cuando se pasa al grupo de edad de 15 a 17 años y una disminución en el riesgo del 4.7 % cuando se pasa al grupo de 18 a 24 años de edad (ver γ_3^*). Análogamente, cuando se pasa al grupo de edad de 25 a 29 años hay una disminución del 1.3 % en el riesgo reportado a través del parámetro γ_4^* y luego hay un repunte

[1]Recordar que los contrastes para el factor sexo son polinomios ortogonales, donde la categoría hombre está asociada al coeficiente $-\frac{2\cos(\pi/4)}{n}$ mientras que las mujeres están asociadas al coeficiente $\frac{2\cos(\pi/4)}{n}$, con $n = 2, ..., 7$.

en el aumento del riesgo de 2.2 % reportado por γ_5^*. Finalmente, el riesgo al pasar al grupo de edad de 60 años o más disminuye en 15.2 % (ver parámetro γ_6^*).

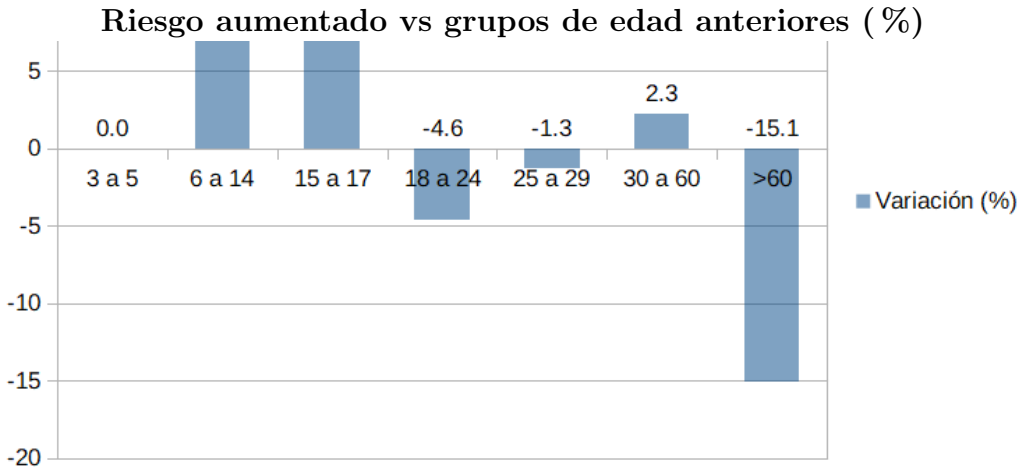


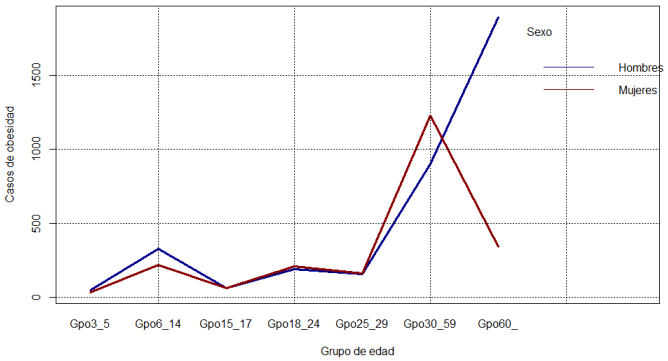
Figura 4.3: Variación en la prevalencia de obesidad al contrastar un grupo de edad específico vs los anteriores para el caso de las mujeres acorde a los valores de los parámetros de interacción reportados en el Cuadro 4.1.

La variación del riesgo relativo de obesidad para las mujeres cuando se salta de un grupo de edad a otro se muestra gráficamente en la Figura 4.3. Esto es consistente con lo que se reportó en la gráfica de interacción de los datos crudos.

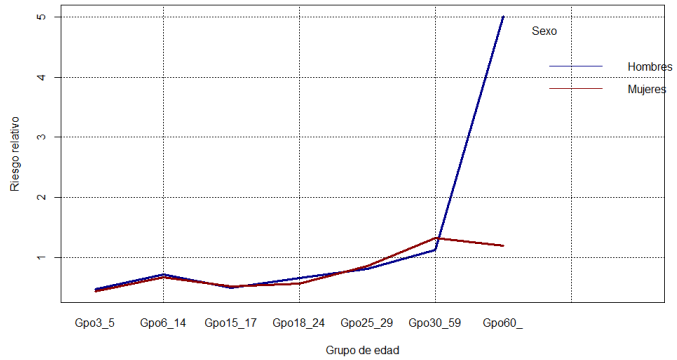
La Figura 4.4 muestra las gráficas de interacción del número de casos observados en la muestra de la ENSANUT 2012 así como del riesgo relativo de la obesidad estimado por el modelo espacial. Debido a la similitud entre las gráfica (a) y (b) de la Figura 4.4, se puede decir que el modelo ajustado es capaz de explicar cualitativamente lo que se observó en los datos, es decir los hombres y las mujeres tienen riesgo igual en edades inferiores a los 15 años.

De acuerdo a la gráfica 4.4 (a), las mujeres tienen más casos de obesidad reportados en la ENSANUT 2012 que los hombres de los 17 a los 60 años mientras que los casos de obesidad para los hombres son mayores que los de las mujeres antes de los 17 años y después de los 59 años.

Estos resultados sugieren que el riesgo de padecer obesidad es similar en edades tempranas y diferentes en edades mayores.



(a) Gráfica de interacción del número de casos de obesidad por hombres y mujeres reportados en la muestra de la ENSANUT 2012.



(b) Gráfica de interacción del riesgo relativo de padecer obesidad estimado por el modelo espacial para hombres y mujeres.

Figura 4.4: Gráficas de interacción del número de casos de obesidad observados (arriba) y del riesgo relativo estimado por el modelo (abajo).

Para personas entre los 25 y 59 años, el modelo sugiere que las mujeres están ligeramente en mayor riesgo que los hombres según la Figura 4.4 lo cual es consistente con lo reportado con otros autores [28]. Sin embargo, para el grupo de edad de 60 años el modelo dice que el riesgo es mucho más grande para hombres que para mujeres. La explicación para estos niveles altos de riesgo es que hubo registros en la ENSANUT 2012 de personas del sexo masculino que representan a muchos hombres a nivel estatal en la muestra[2], entonces al dividir entre el número esperado de casos de varones de 60 años o más por

[2]El diseño de muestra es tal que una persona encuestada tiene representación a nivel estatal por lo que las personas representadas por esta persona pueden o no vivir en el municipio del encuestado.

municipio (que puede ser pequeño) los cocientes se disparan.

Mapas de riesgo

Una vez que los parámetros del modelo 3.2 fueron estimados es posible graficar los valores estimados del riesgo relativo para cada municipio. Recordando que el modelo planteado es:

$$Y_i \sim Poi(c_i x_i)$$

donde x_i es riesgo relativo entonces basta con sustituir los parámetros estimados en la ecuación (3.2) y tomar los valores exponenciados, donde los factores *sexo* y *grupo de edad* tienen asociadas las matrices de contraste C_a y C_b descritas anteriormente así como las matrices *dummy* $\boldsymbol{X}_a$ y $\boldsymbol{X}_b$ acorde a la notación de la expresión (2.33).

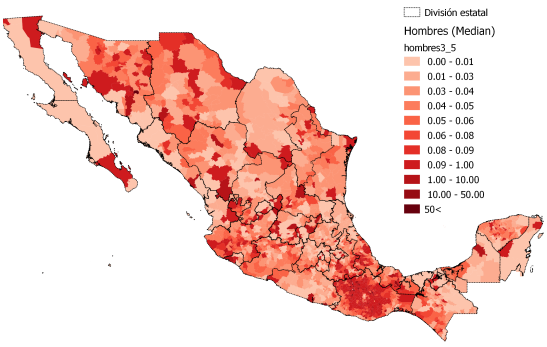
Las figuras 4.5 y 4.6 muestran el riesgo relativo estimado por sexo y grupo de edad tomando las medianas de los parámetros que están reportadas en el cuadro 4.1. Para el caso de los hombres, se puede ver en estas figuras que la progresión del riesgo relativo va aumentando conforme se pasa de un grupo de edad a otro donde el registro más alto se obtiene con el grupo de 60 años o más.

Para las mujeres, el riesgo relativo también es creciente conforme se pasa de un grupo de edad a otro con quizás la excepción del grupo de 60 años y más ya que en éste hay una ligera disminución respecto al grupo de 30 a 59 años.

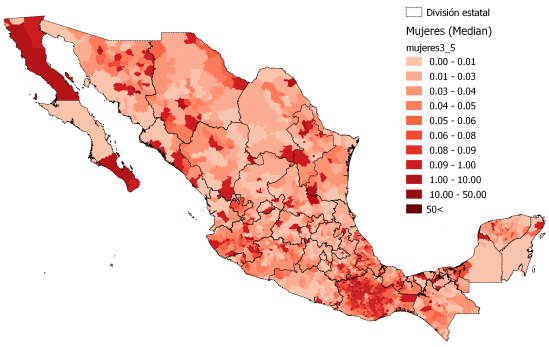
El análisis visual de los mapas es consistente tanto para hombres como mujeres con lo que se reporta en la gráfica de interacción 4.4 donde se comparan los riesgos relativos agrupados por sexo y grupo de edad. Es decir, las tendencias de los mapas se ven reflejadas en el gráfico de interacción y viceversa.

En las figuras 4.7 y 4.8 se observa un conglomerado de municipios con alto riesgo de riesgo que corresponden a los estados de Oaxaca y Chiapas principalmente. Este conglomerado es muy similar al conglomerado que aparece en el mapa del índice de desarrollo humano reportado en la figura 3.3.

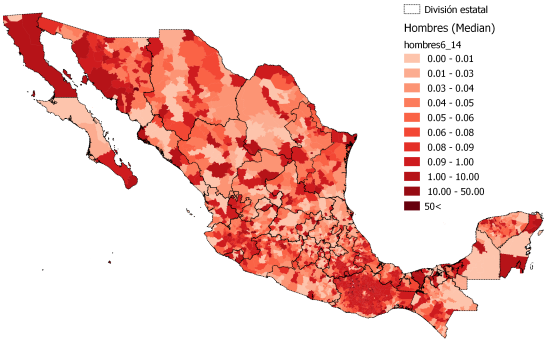
Por otro lado, en un estudio hecho por Barquera [28] se reporta que en el periodo del año 2000 al 2012 los 10 estados con mayor prevalencia de obesidad para hombres y mujeres mayores de 20 años fueron Chiapas, Oaxaca, Hidalgo, San Luis Potosí, Quintana Roo, Puebla, Guerrero, Tlaxcala, Ciudad de México y Guanajuato. En ese reporte de Barquera[28] aparece la Figura 4.9 y como se puede apreciar hay consistencia en el sentido de que Chiapas y Oaxaca encabezan los estados con mayor prevalencia de obesidad.



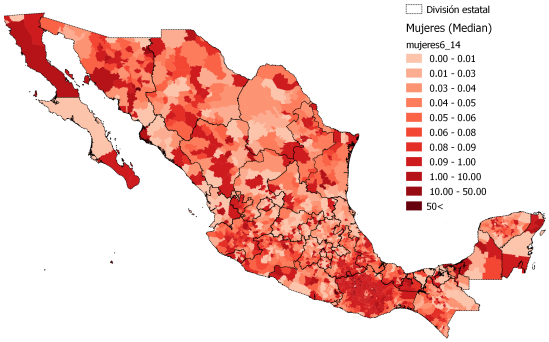
(a) Riesgo hombres 3 a 5 años



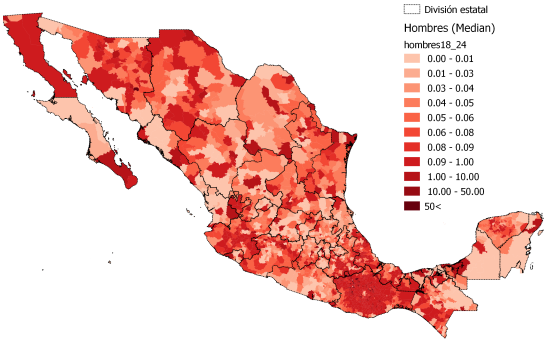
(b) Riesgo mujeres 3 a 5 años



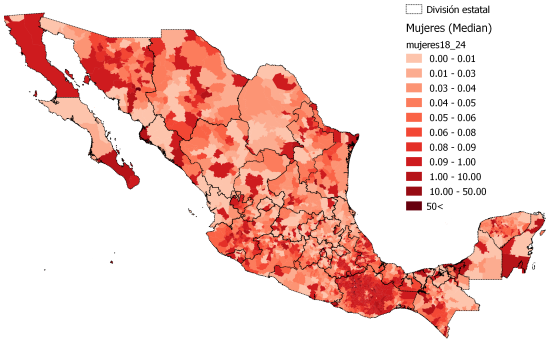
(c) Riesgo hombres 6 a 14 años



(d) Riesgo mujeres 6 a 14 años

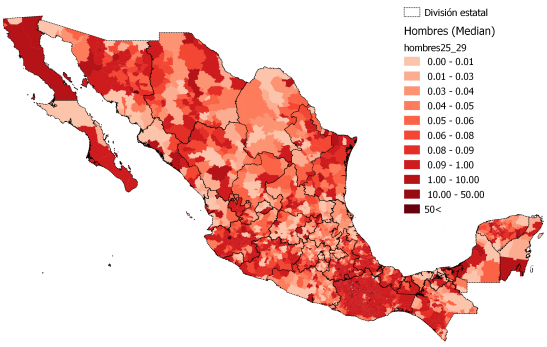


(e) Riesgo hombres 18 a 24 años

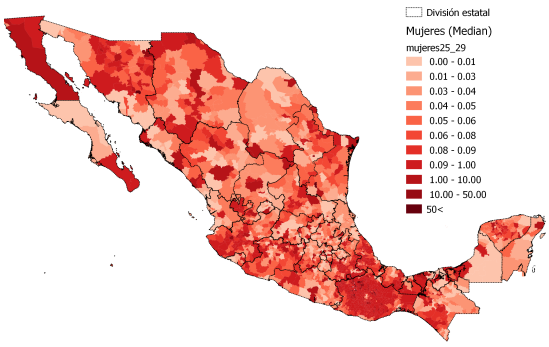


(f) Riesgo mujeres 18 a 24 años

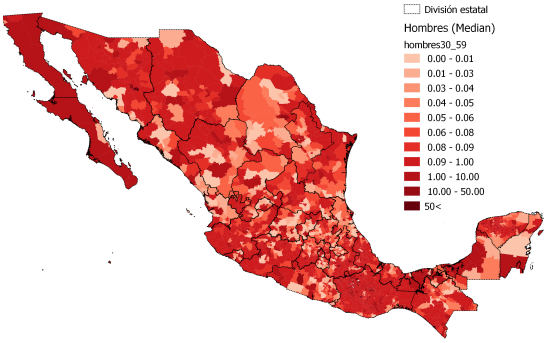
Figura 4.5: Mapas de la mediana del riesgo relativo por grupo de edad y sexo.



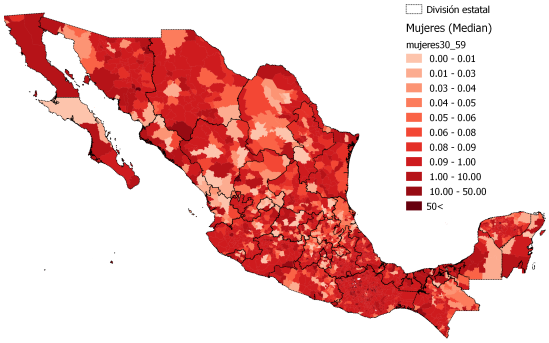
(a) Riesgo hombres 25 a 29 años



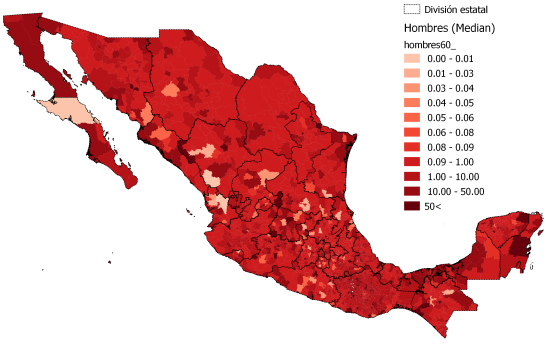
(b) Riesgo mujeres 25 a 29 años



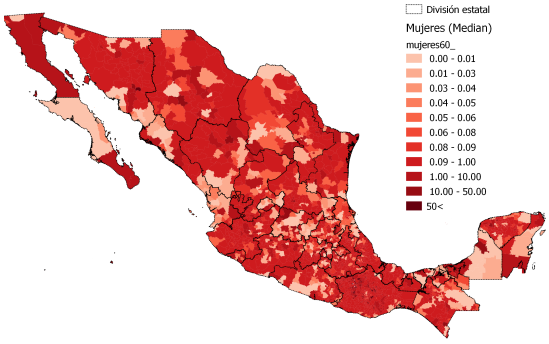
(c) Riesgo hombres 30 a 59 años



(d) Riesgo mujeres 30 a 59 años

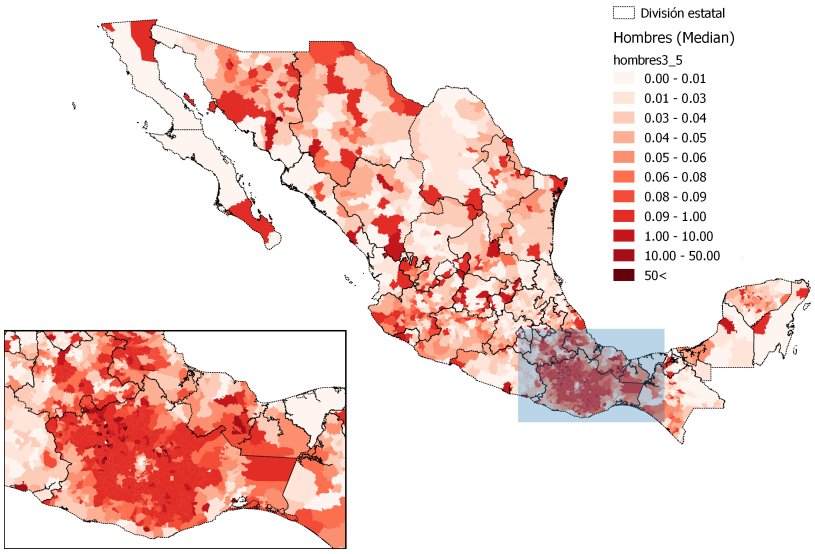


(e) Riesgo hombres 60 años o más

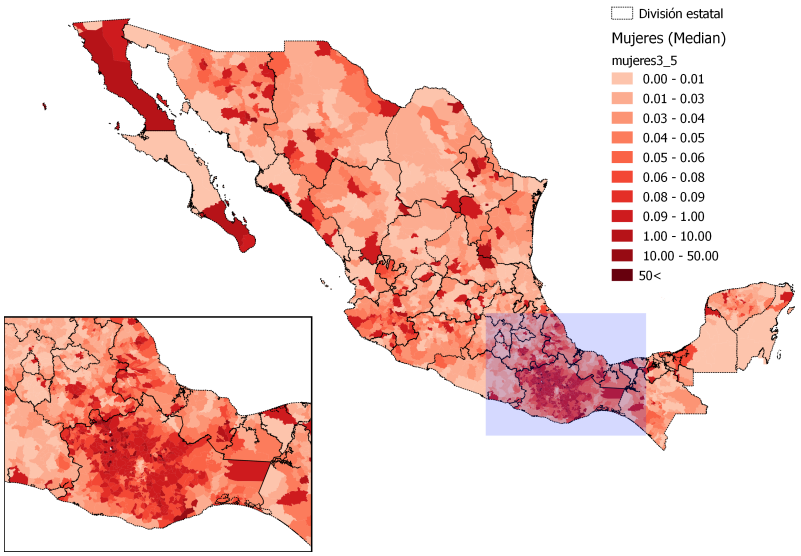


(f) Riesgo mujeres 60 años o más

Figura 4.6: Mapas de la mediana del riesgo relativo por grupo de edad y sexo.

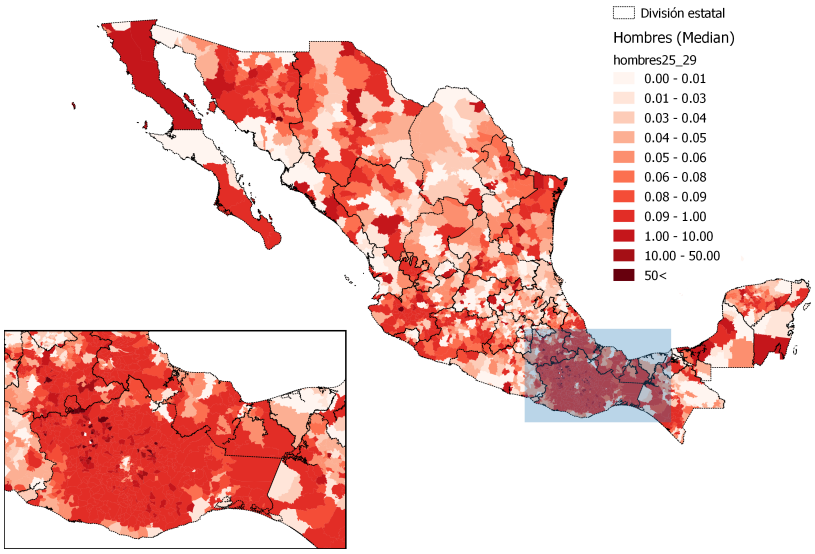


(a) Riesgo hombres 3 a 5 años

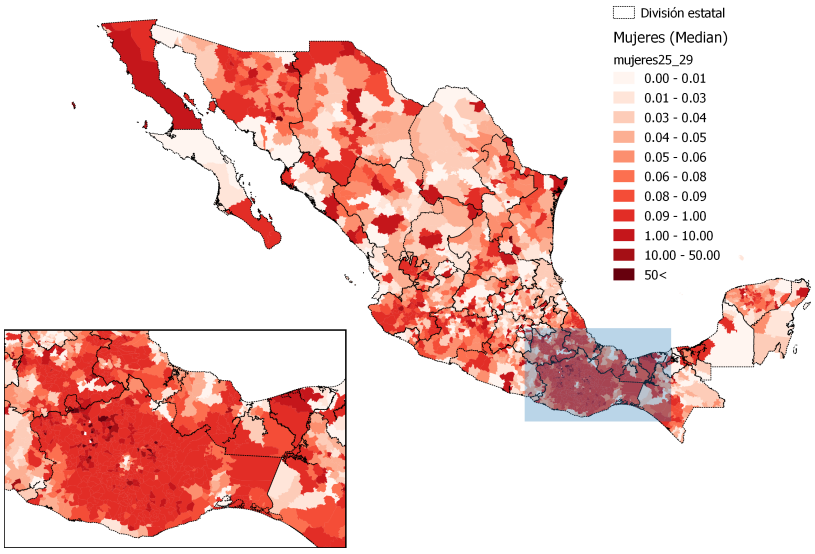


(b) Riesgo mujeres 3 a 5 años

Figura 4.7: Conglomerado de riesgo relativo en la zona de Oaxaca



(a) Riesgo hombres 25 a 29 años



(b) Riesgo mujeres 25 a 29 años

Figura 4.8: Conglomerado de riesgo relativo en la zona de Oaxaca

Cabe mencionar que los resultados generales de la ENSANUT 2006 reportados también por Barquera [29] muestran que los grupos de edad con mayor prevalencia de obesidad independientemente del sexo fueron los grupos de 50 a 59 años seguido del de 40 a 49 años de edad. Esto es similar con los resultados obtenidos en este trabajo para el grupo de mujeres acorde a la gráfica de interacción. El estudio de la ENSANUT 2006 señala que las mujeres tienen una prevalencia de 35.4 % mientras que para los hombres es de 24 % para el año 2006. A pesar de que este resultado no es comparable de manera general con el análisis presentado aquí puesto el modelo espacial incluyó como factor de entrada a los grupos de edad, modelo de Poisson señala que el riesgo de las mujeres es mayor en edades de los 25 a 59 años y esto coincide a groso modo con la ENSANUT 2006.

El Cuadro 4.4 contiene las prevalencias de obesidad para hombres y mujeres de 20 años o más reportadas acorde a los datos de la ENSANUT 2012 por Barquera. En general, la prevalencia es mayor para las mujeres que para los hombres excepto en la categoría de obesidad abdominal donde existe una diferencia de aproximadamente 18 puntos porcentuales. Es importante señalar que en la ENSANUT 2012, el criterio para determinar si una persona tiene obesidad abdominal está basado en si la longitud del diámetro del abdomen excede los 80 y 90 centímetros para mujeres y hombres respectivamente mientras que nuestro estudio el sólo consideramos si una persona encuestada tiene o no obesidad basados en el índice de masa corporal. La ENSANUT en su versión de 2006 y 2012 sugiere que el grupo de edad con mayor prevalencia es el que abarca de los 30 a 60 años de edad lo cual es consistente con las gráficas de interacción de la Figura 4.4.

En otro estudio de Barquera[30] publicado en 2003, el autor hace un análisis de las tasas de mortalidad a nivel geográfico por grupo de edad de las defunciones por diabetes mellitus entre los años 1980 y 2000. En ese trabajo el autor indica que hay un incremento en las tendencias de esas tasas de mortalidad por grupo de edad y sexo además señala que hay diferencias notables en la distribución geográfica las tasas de mortalidad de la diabetes mellitus siendo la región del sur la que experimento un mayor cambio (un 92 % del año 2000 respecto a 1980). Ese estudio también señala que los estados con mayor tasa de mortalidad estandarizada de diabetes mellitus son Coahuila seguido por Baja California Norte [3] y Aguascalientes. Sin embargo, el estado de Oaxaca aparece en el último lugar con una tasa de mortalidad estandarizada de

[3]Baja California Norte tiene conglomerados de municipios con riesgo notable para casi todos los grupos de edad y sexo acorde a la Figuras4.5 y 4.6

diabetes mellitus de 0.65 durante el periodo de 1980 al año 2000. Esto se debe a que Barquera analiza un periodo de 20 años para el cálculo de las tasas de mortalidad estandarizadas de diabetes por estado, sin embargo el autor reconoce que la región de México que tuvieron un mayor incremento en las tasas de mortalidad por diabetes en el año 2000 relativo a 1980 fue el sur de México con un 928 %.

En el análisis espacial que se presenta en este trabajo, se puede notar que la zona correspondiente al estado de Oaxaca junto con algunas partes de Veracruz, Chiapas y Guerrero destaca por tener un conglomerado de municipios con un riesgo relativo más alto que en otras regiones de México para el año 2012 mientras que su tasa de mortalidad de diabetes mellitus es de 0.65 en el año 2000. La diabetes mellitus y la obesidad son enfermedades muy relacionadas y en particular para el estado de Oaxaca los resultados del modelo espacial Poisson muestran semejanza con lo que el aumento de 92 % en la tasas de mortalidad de diabetes de 1980 contrastada con la tasa del año 2000 para este estado.

Cuadro 4.4: Prevalencia de obesidad reportada por sexo y nivel de obesidad para personas de 20 años o más en escala porcentual[28] según datos de le ENSANUT 2012.

Sexo/Nivel	Normal	Sobrepeso	Obesidad I	Obesidad II	Obesidad III	Obesidad abdominal
Mujeres	25.6	35.5	24.0	9.4	4.1	64.5
Hombres	29.6	42.6	20.1	5.0	1.8	82.8

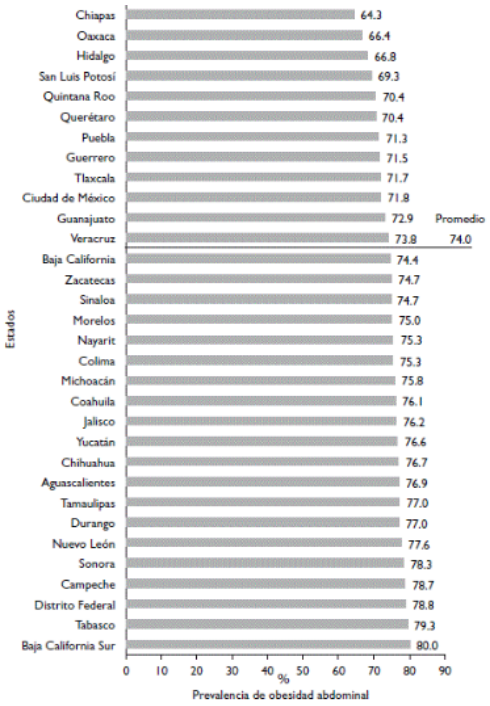


Figura 4.9: Prevalencia estatal de obesidad abdominal en adultos mayores de 20 años reportado por Barquera [28] en los años 2000-2012. Datos ajustados por el diseño complejo de la muestra. Mujeres embarazadas fueron excluidas del análisis.

4.2.2. Prueba I de Moran

Cómo se mencionó anteriormente el análisis de residuales es importante para saber si el al ajuste del modelo estadístico fue razonable. La prueba de Moran aplicada sobre los residuales[4] del modelo ayuda a saber si el modelo fue capaz de recuperar la estructura espacial. Es decir, en el caso donde la estructura espacial está todavía presente en los residuales nos dice que las covariables en conjunto con el campo markoviano no pudieron explicar ni recuperar la variación sobre la variable respuesta (en este caso el modelo no recuperaría el riesgo relativo de obesidad a nivel municipal).

En el escenario donde los residuales continúan preservando la estructura espacial, la prueba de Moran reportará un valor p "grande"de otra manera éste será "pequeño" cuando los residuales tienen una patrón aleatorio que por ende no refleja estructura espacial alguna. En esencia la prueba de hipótesis es:

[4]Recordar que los residuales se calculan $r_i = Y_i - \hat{Y}_i$.

$H_0 :=$ No hay correlación espacial vs $H_1 :=$ Hay correlación espacial

Los valores de la prueba I de Moran por grupo de edad y sexo se reportan en el cuadro 4.5). Como se puede ver, la mayoría de los valores p son mayores que 0.1 y entonces se concluye que el modelo propuesto logra capturar la variación espacial con las covariables empleadas que son IDH, tasa de muerte por enfermedades cardiovasculares, edad y sexo. Los casos donde el modelo tiene dificultades en el ajuste son: mujeres de 3 a 5 años y hombres de 6 a 14 años. El gráfico 4.10 muestra los residuales del modelo restringido a las mujeres

Cuadro 4.5: Resumen estadístico de la prueba de Moran agrupada por grupo de edad y sexo sobre los residuales del modelo.

Sexo y Grupo de edad	Estadístico I	$E(I)$	p-valor
3 a 5 años (hombres)	-0.0226	-0.0014	0.780
3 a 5 años (mujeres)	0.0576	-0.0014	0.016
6 a 14 años (hombres)	0.0419	-0.0014	0.057
6 a 14 años (mujeres)	-0.0239	-0.0014	0.785
15 a 17 años (hombres)	0.0151	-0.0014	0.269
15 a 17 años (mujeres)	-0.0129	-0.0014	0.655
18 a 24 años (hombres)	-0.0366	-0.0014	0.888
18 a 24 años (mujeres)	-0.0037	-0.0014	0.538
25 a 29 años (hombres)	0.0234	-0.0014	0.145
25 a 29 años (mujeres)	0.0036	-0.0014	0.385
30 a 59 años (hombres)	0.0058	-0.0014	0.394
30 a 59 años (mujeres)	-0.0042	-0.0014	0.545
60 años o más (hombres)	-0.0004	-0.0014	0.486
60 años o más (mujeres)	-0.0058	-0.0014	0.564

entre 3 y 5 años de edad. El valor p reportado por la prueba de Moran acorde a la tabla 4.5 es 0.016 lo cual rechaza la hipótesis de que no hay correlación espacial y en consecuencia decimos que el modelo no fue capaz de explicar la variabilidad del logaritmo del riesgo relativo en el caso de mujeres entre 3 y 5 años de edad. Una conclusión similar aplica para el caso de los hombres en el grupo de edad de 6 a 14 años donde el valor p reportado por la tabla 4.5 fue 0.057.

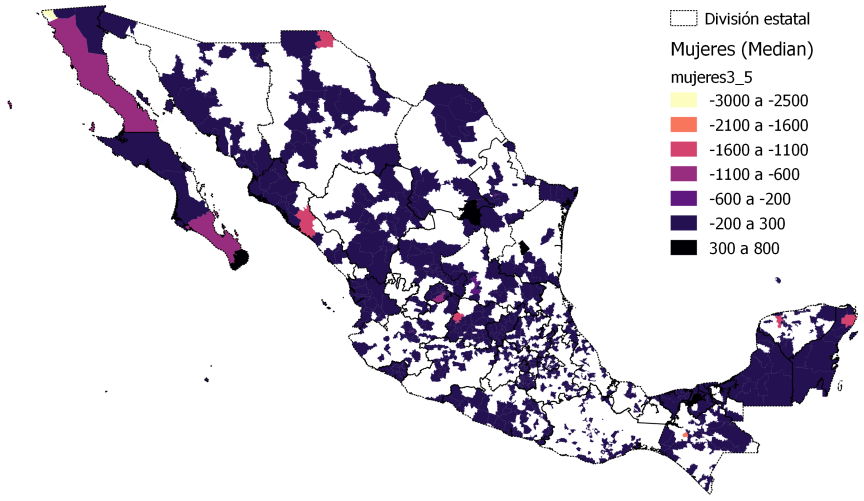


Figura 4.10: Residuales del modelo del grupo mujeres entre 3 y 5 años de edad. El valor p fue 0.016 **lo cual sugiere que hay un patrón espacial presente.**

4.2.3. Selección de Modelo

Un experimento que se hizo para la selección de modelos fue ajustar un modelo estadístico que incluyera únicamente al intercepto para luego ser comparado contra el modelo espacial Poisson. Los resultados son los siguientes:

1. Modelo 1 (intercepto, covariables, efecto espacial y ruido): $x_i = \mu + \mu_1 IDH + Edad + Sexo + Edad \cdot Sexo + u_i + v_i$, el criterio de información de devianza es $CID = 3,707,712$

2. Modelo 2(sólo intercepto): $x_i = \mu$, el criterio de información de devianza es $CID = 5,374,947$

lo anterior sugiere que el modelo 1 (el cual tome en cuenta el intercepto, covariables, efecto espacial y ruido) es mejor que el modelo 2 en términos de mejor ajuste y complejidad ya que la devianza del modelo 1 es menor que la del modelo 2.

Capítulo 5

CONCLUSIONES

La base de este trabajo es la Encuesta Nacional de Salud (ENSANUT) del 2012. Un factor que influye directamente en el modelo espacial es el diseño y aplicación de la encuesta. La unidad básica muestral de la ENSANUT es el Área Geoestadística Básica (AGEB) la cual no necesariamente coincide con un municipio (y mucho menos un estado). El reto fue en que se deseaba construir un modelo que explicará el riesgo relativo de padecer obesidad a un nivel granular de municipios. Después de un proceso de limpieza de datos e incluir las tasas de mortalidad por diabetes a nivel municipal para ajustar el muestreo, se pudo estimar el número de casos de obesidad por grupo de edad y sexo a nivel municipal y entonces pasar a la parte de modelación.

El método presentado en este trabajo es un primer intento de sacar provecho a los datos geo-referenciados de la ENSANUT en su versión del año 2012 que no han sido estudiados a un nivel de municipios para apoyar la toma de decisiones que conciernen al diagnóstico, diseño y ejecución de estrategias de prevención de sobrepeso y obesidad.

El análisis de los datos requirió un almacenamiento y procesamiento óptimo ya que se simularon cadenas de Markov para los parámetros del modelo así como para cada uno de los 2456 municipios por lo que el tiempo de simulación puede ser tardado en especial cuando se incluyen variables categóricas como el sexo y grupo de edad.

A pesar de que existen versiones "mejoradas"de modelo de Besag y York, en este trabajo se utilizó el modelo original y éste logró recuperar el patrón espacial con las covariables empleadas basado en lo reportado por la prueba de Moran. Además, el modelo propuesto incluye variables categóricas y no hay mucha información en la literatura referente a la estadística espacial donde se incluyan variables categóricas.

El modelo propuesto logró recuperar de manera cualitativa lo reportado por la ENSANUT 2012, es decir los conteos del número casos de obesidad por grupo y sexo es parecido a los riesgos estimados por edad y sexo. Basado en las estimaciones del modelo, la conclusión es que tanto hombres como mujeres tienen riesgos muy parecidos en edades anteriores a los 17 años; las mujeres tienen más riesgo en edades entre los 25 a 59 años y finamente los hombres tienen mucho más riesgo en edades mayores a los 59 años contrario a las mujeres que muestran una importante disminución en el riesgo.

Los resultados del modelo espacial Poisson muestran que las regiones de Oaxaca y Chiapas mayormente tienen alto riesgo de obesidad para todos los grupos de edad y sexo. Las entidades de Jalisco, Durango y Michoacán tienen un conglomerado de municipios con un riesgo relativo notorio aunque no es tan marcado como el de Oaxaca y Chiapas.

De manera general, los resultados obtenidos en los mapas sugieren que las regiones con mayor riesgo son parecidas a las ya reportadas por otros autores sin embargo en este trabajo se segregó por grupos de edad y sexo por municipio lo cual hace novedoso a este estudio en la literatura. De hecho la ENSANUT 2006 reportó que las mujeres tienen mayor riesgo que los hombres pero como se acaba de ver existen claroscuros en el riesgo dependiendo los grupos de edad y sexo.

Este trabajo encontró que las mujeres tienen ligeramente mayor riesgo de padecer obesidad que los hombres en edades que comprenden de los 30 a 59 años siendo los estados de Oaxaca, Chiapas, Jalisco, Baja California Norte , Chichuachua y Durango los que destacan. Esta tendencia se revierte después de los 60 años ya que el modelo sugiere que los hombres están en mucho mayor riesgo que las mujeres.

Resultados de estudios previos son similares con los obtenidos aquí en el caso del estado de Oaxaca. Esos estudios señalan a Oaxaca como la entidad federativa que tuvo el incremento porcentual más grande en la tasa de diabetes mellitus del año 2000 comparada contra el año de 1980. El hecho de que Oaxaca encabece la lista en tasas de obesidad y con un incremento alto en la tasas de diabetes mellitus sugiere que algo ocurre en esa zona. Desde el punto de vista del modelo estadístico el valor alto en el riesgo relativo es explicado debido a que Oaxaca es un estado con carencias en el sistema de salud y una economía no tan desarrollada lo cual se refleja en el índice de desarrollo humano (IDH). El formular otra explicación a partir de los datos de la ENSANUT 2012 sería aventurado.

Es importante señalar que a nivel internacional son muy pocos los estudios donde se haga un análisis del sobrepeso y obesidad por grupos de edad y sexo. Particularmente en América Latina no se encontraron estudios similares por lo que es complicado extrapolar conclusiones de este trabajo hacia Latinoamérica.

APÉNDICE

Descripción de las covariables reportadas en la ENSANUT 2012 codificadas en su versión original.

1. mun.mun-alimento %: Porcentaje de personas con rezago nutricional por municipio.
2. mun.mun-servsalud %: Porcentaje de personas sin derechohabiencia por municipio.
3. mun.mun-ingpercap: Ingreso per capita por municipio.
4. mun.mun-indice-DH: Índice de desarrollo humano por municipio.
5. mun.mun-inginf %: Porcentaje de personas que reciben ingresos inferiores a la línea media de bienestar por municipio.
6. mun.mun-rezeduc %: Porcentaje de personas con rezago educativo por municipio.
7. mun.mun-indice-RN: Índice de rezago nutricional por municipio.
8. mun.rs.IRS10: Índice de rezago social por municipio.
9. mun.marg.in-marg: Índice de marginación por municipio.

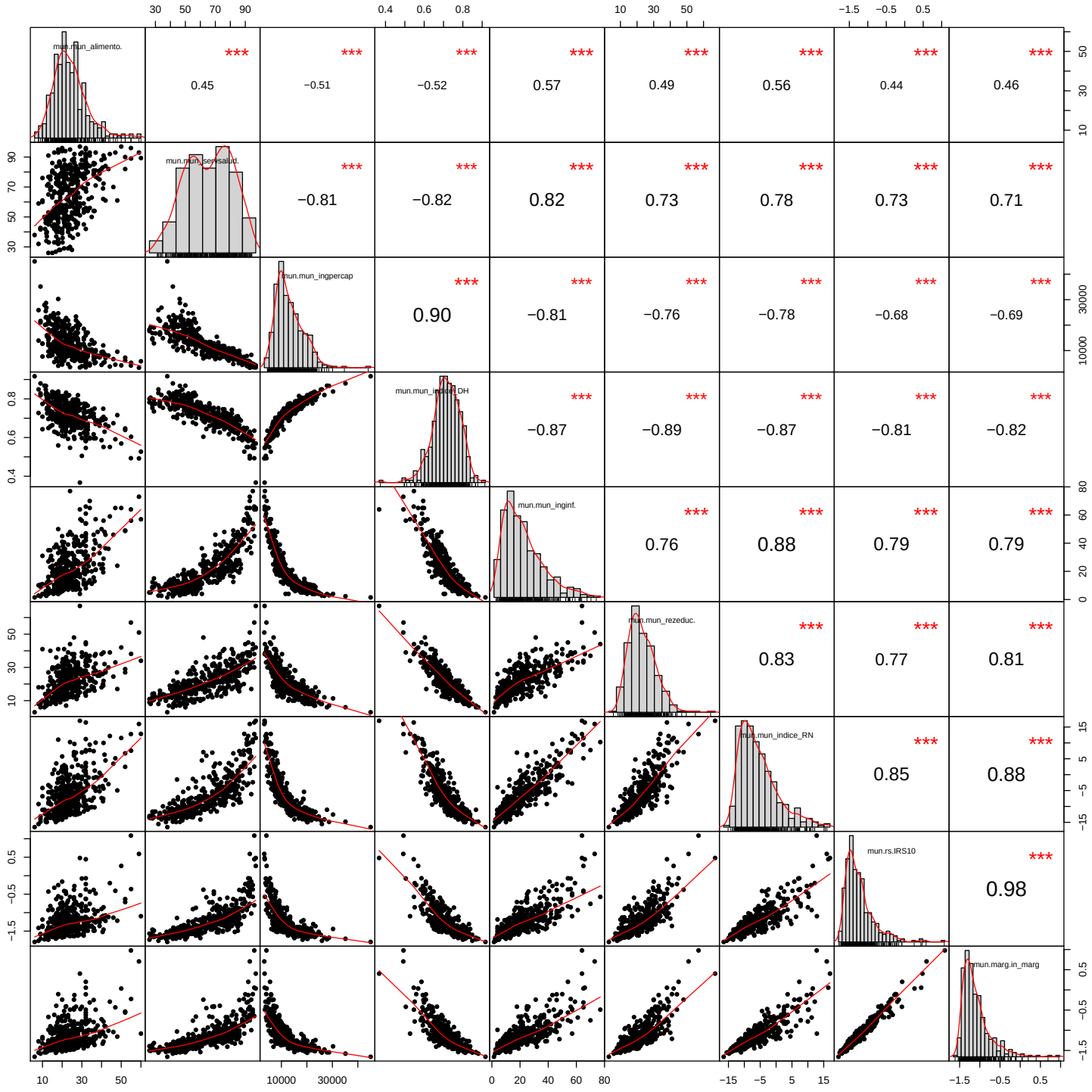


Figura 5.1: Gráfica de dispersión de las covariables socioeconómicas a nivel municipal reportadas en la ENSANUT 2012.

Bibliografía

[1] "Los Impuestos a los Refrescos y a las Bebidas Azucaradas como Medida de Salud Pública." https://www.paho.org/mex/index.php?option=com_content&view=article&id=627:los-impuestos-refrescos-bebidas-azucaradas-medida-salud-publica&Itemid=499. Accesado: 2019-04-30.

[2] "Obesidad y sobrepeso." https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight. Accesado: 2019-04-30.

[3] "Ocde (2011), health at a glance 2011: Ocde indicators, ocde publishing.,"

[4] B. A. Swinburn, G. Sacks, K. D. Hall, K. McPherson, D. T. Finegood, M. L. Moodie, and S. L. Gortmaker, "The global obesity pandemic: shaped by global drivers and local environments," *The Lancet*, vol. 378, no. 9793, pp. 804–814, 2011.

[5] "The burden of disease project, 2010. [www document].,"

[6] K. Rtveladze, T. Marsh, S. Barquera, L. M. S. Romero, D. Levy, G. Melendez, L. Webber, F. Kilpi, K. McPherson, and M. Brown, "Obesity prevalence in mexico: impact on health and economic burden," *Public health nutrition*, vol. 17, no. 1, pp. 233–239, 2014.

[7] "Estadísticas de la ocde sobre la salud 2014 méxico en comparación."

[8] M. Romero-Martínez, T. Shamah-Levy, A. Franco-Núñez, S. Villalpando, L. Cuevas-Nasu, J. P. Gutiérrez, and J. Á. Rivera-Dommarco, "Encuesta nacional de salud y nutrición 2012: diseño y cobertura," *salud pública de méxico*, vol. 55, pp. S332–S340, 2013.

[9] N. Cressie, *Statistics for Spatial Data*. Wiley Series in Probability and Statistics, Wiley, 2015.

[10] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 192–225, 1974.

[11] S. Z. Li, *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[12] H. Rue and L. Held, *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[13] J. Besag, J. York, and A. Mollié, "Bayesian image restoration, with two applications in spatial statistics," *Annals of the institute of statistical mathematics*, vol. 43, no. 1, pp. 1–20, 1991.

[14] D. Lee, "Carbayes: An r package for bayesian spatial modeling with conditional autoregressive priors," *Journal of Statistical Software, Articles*, vol. 55, no. 13, pp. 1–24, 2013.

[15] F. Gerber, R. Furrer, *et al.*, "Pitfalls in the implementation of bayesian hierarchical modeling of areal count data: An illustration using bym and leroux models," *Journal of Statistical Software, Code Snippets*, vol. 63, no. 1, pp. 1–32, 2015.

[16] W. N. Venables and B. D. Ripley, *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[17] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[18] J. M. Bernardo, "The concept of exchangeability and its applications,"

[19] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde, "Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models," tech. rep., Research Report, 98-009, 1998.

[20] A. Dempster, "The direct use of likelihood for significance testing," in *Proceedings of Conference on Foundational Questions in Statistical Inference, Aarhus, May 7-12, 1973*, no. 1, p. 335, Department of Theoretical Statistics, Institute of Mathematics, University, 1974.

[21] "Sistema Nacional de Informaciǿn Municipal." `http://www.snim.rami.gob.mx/`. Accesado: 2019-08-10.

[22] B. G. Leroux, X. Lei, and N. Breslow, "Estimation of disease rates in small areas: a new mixed model for spatial dependence," in *Statistical models in epidemiology, the environment, and clinical trials*, pp. 179–191, Springer, 2000.

[23] D. Lee and R. Mitchell, "Boundary detection in disease mapping studies," *Biostatistics*, vol. 13, no. 3, pp. 415–426, 2012.

[24] D. Lee and C. Sarran, "Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies," *Environmetrics*, vol. 26, no. 7, pp. 477–487, 2015.

[25] L. McLaren, "Socioeconomic status and obesity," *Epidemiologic reviews*, vol. 29, no. 1, pp. 29–48, 2007.

[26] A. A. Brewis, *Obesity: Cultural and biocultural perspectives*. Rutgers University Press, 2010.

[27] Z. Khazaei, M. Sohrabivafa, I. Darvishi, H. Naemi, and E. Goodarzi, "Relation between obesity prevalence and the human development index and its components: an updated study on the asian population," *Journal of Public Health*, pp. 1–7, 2020.

[28] S. Barquera, I. Campos-Nonato, L. Hernández-Barrera, A. Pedroza, and J. A. Rivera-Dommarco, "Prevalencia de obesidad en adultos mexicanos, 2000-2012," *Salud pública de México*, vol. 55, pp. S151–S160, 2013.

[29] S. B. Cervera, I. Campos-Nonato, R. Rojas, and J. Rivera, "Obesidad en méxico: epidemiología y políticas de salud para su control y prevención," *Gaceta Médica de México*, vol. 146, no. 6, pp. 397–407, 2010.

[30] S. Barquera, V. Tovar-Guzmán, I. Campos-Nonato, C. González-Villalpando, and J. Rivera-Dommarco, "Geography of diabetes mellitus mortality in mexico: an epidemiologic transition analysis," *Archives of medical research*, vol. 34, no. 5, pp. 407–414, 2003.