UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Implementación de Modelo Multivariable de Regresión Lineal con enfoque Bayesiano

T E S I S

QUE PARA OBTENER EL TÍTULO DE**:**

Licenciado en Ciencias de la Tierra

P R E S E N T A :

Eder Luis Salazar Díaz

DIRECTOR DE TESIS:
Dr. Adolfo Magaldi Hermosillo

Ciudad Universitaria, CDMX 2019

Hoja de datos del jurado

1. Datos del alumno
   Eder Luis Salazar Díaz
   Licenciatura en Ciencias de la Tierra
   Facultad de Ciencias
   Universidad Nacional Autónoma de México
   309313427

2. Datos del tutor
   Dr. Dr. Adolfo Magaldi Hermosillo
   Departamento de Instrumentación y Observación Atmosférica
   Centro de Ciencias de la Atmósfera
   Universidad Nacional Autónoma de México

3. Datos del sinodal 1
   Dra. Dara Salcedo Golzález
   Unidad Multidisciplinaria de Docencia e Investigación
   Facultad de Ciencias
   Universidad Nacional Autónoma de México

4. Datos del sinodal 2
   Dr. Harry Alvarez Ospina
   Departamento de Física
   Facultad de Ciencias
   Universidad Nacional Autónoma de México

5. Datos del sinodal 3
   Dr. Gerardo Hernández Dueñas
   Investigador Asociado C
   Instituto de Matemáticas
   Universidad Nacional Autónoma de México

6. Datos del sinodal 4
   Dr. Iván Yassmany Hernández Paniagua
   Catedrático CONACYT
   CentroMet
   Consorcio para el Estudio de Zonas Metropolitanas

Datos del trabajo escrito
Implementación de Modelo Multivariable de Regresión Lineal con enfoque Bayesiano
66 pp.
2019

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

A la RAMA y RUOA por facilitarme los datos con que trabajar.

A la UNAM, mi segunda casa. Es para mi un orgullo pertenecer a la máxima casa de estudios del país. Le agradezco la oportunidad académica y deportiva que me ha dado tantos años de felicidad.

A Adolfo por la libertad y paciencia, que me enseño a hacerlo por gusto y no por compromiso.

A Dara que me puso un pie en las Ciencias Atmosféricas.

A mis sinodales, sin su observaciones mi tesis estaría perdida.

A cada miembro de mi familia, por enseñarme a subir escalando viendo de vez en cuando para abajo. Especialmente a Olivia y Luis, madre y padre, gracias por el amor incondicional que nos han dado a mi hermano y a mi. Sus logros son nuestros logros, mis logros son por ustedes.

A Mayan por estar en mi vida.

Y a mis amigos y amigas, por las horas de ocio, del té y las tantas horas que pasamos platicando.

A todos, muchas gracias por compartirme un poco de ustedes.

Resumen

Las grandes ciudades presentan cada vez mayores concentraciones de contaminantes, lo que pone en riesgo la salud. Por lo tanto, surge la necesidad de buscar herramientas que brinden información en tiempo real de la exposición a los contaminantes, por parte de la población y que ayuden en la toma de decisiones. Esto motiva la elaboración y construcción de un modelo basado en la estadística Bayesiana, que represente concentraciones de PM a partir de: la cantidad y tipo de vehículos que circulan sobre la avenida Delfín Madrigal (a un costado de metro Universidad), así como distintas variables meteorológicas. Este trabajo tiene la finalidad de obtener un modelo que permita reconstruir el comportamiento de las concentraciones de material particulado, para usarlo como base en el pronóstico de dicho contaminante. El modelo con enfoque Bayesiano que se obtuvo, se ajusta muy bien a las concentraciones observadas de PM_{10} de los días 25, 26 y 27 del mes de marzo del 2015.

Capítulo 1

Introducción

La contaminación del aire representa una gran amenaza para la población (OMS, 2016), debido a la exposición de los contaminantes. En particular; el dióxido de azufre (SO_2), el dióxido de nitrógeno (NO_2), el monóxido de carbono (CO), el material particulado (PM) y el ozono (O_3), a los que también se les llama contaminantes criterio, por el riesgo que representa a la salud (Chen et al., 2007), por lo que es importante conocer los factores involucrados en su transporte y emisión sobre las grandes ciudades. En la tabla 1.1 se presentan los límites de exposición recomendados por la OMS en el 2016 (OMS, 2016).

El aerosol atmosférico o Material Particulado (PM, por sus siglas en ingles) se definen técnicamente a la mezcla de partículas solidas y liquidas que sean los suficientemente pequeñas para encontrarse suspendidas en la atmósfera. Si este tiene un tamaño igual o menor a 10 micras llega a ser lo suficientemente pequeño para atravesar los mecanismos naturales de defensas pulmonares (Ahrens, 2012), agudizando problemas respiratorios y cardiovasculares (Pearce and Crowards, 1996). Por lo que en este trabajo se seleccionó este contaminante como objeto de estudio.

Conocer las condiciones meteorológicas de cualquier lugar es fundamental para comprender la formación, transformación, difusión, transporte y eliminación de los contaminantes atmosféricos, que son emitidos y transportados directamente o formados a la atmósfera. Al incluir las variables meteorológicas podemos mejorar los modelos atmosféricos para proporcionar mejores pronósticos de calidad del aire. Esto representa un desafío importante, no trivial, por lo que se necesario el uso de estadística que pueda trabajar con diversas variables.

La estadística Bayesiana representa una alternativa a la estadística convencional. Surge a

Directrices según la OMS, 2005	
PM_{10}	20 $\mu g/m^3$ de media anual
	50 $\mu g/m^3$ de media en 24h
O_3	100 $\mu g/m^3$ de media en 8h
NO_2	40 $\mu g/m^3$ de media anual
	200 $\mu g/m^3$ de media en 1h
SO_2	20 $\mu g/m^3$ de media en 24h
	500 $\mu g/m^3$ de media en 10 min

Tabla 1.1: Limites de exposición de los contaminantes criterio (OMS, 2016).

mediados del siglo XVIII y creada por Thomas Bayes, quien introduce el teorema de Bayes que representa el uso de probabilidades condicionales, donde un evento subsecuente a otro da información de un evento previo, siendo este la base de todo un enfoque estadístico. Aunque fue abandonado su uso casi en su totalidad en los siglos posteriores a su creación, en la actualidad ha resurgido este teorema y ha sido ampliamente aplicado en los distintos campos de la ciencia, ciencias de la salud y ciencias sociales (Gutiérrez Peña, 2013). Sin embargo aún se mantiene el debate, entre los enfoques de la estadística clásica y Bayesiana.

La estadística Bayesiana permite el uso de información *a priori*, en base a nuestro conocimiento previo a algún evento. También usa los parámetros, características numéricas de la población, como aleatorias con lo que podemos obtener grados de " creencia "sobre las variables, así no es necesario el uso de muestras grandes. Debido a estas y otras ventajas en la visión Bayesiana, en este trabajo se realiza un modelo con base en la estadística Bayesiana, que reproduzca y se comporte de acuerdo a las concentraciones del material particulado observados en la estación del Centro de Ciencias de la Atmósfera(CCA) de la Red Automática de Monitoreo Atmosférico (RAMA) (Autores, 2017). Esto con la finalidad de realizar en un futuro el pronóstico de dicho contaminante. Tomando como variables las fuentes móviles observadas en la avenida de Delfín Madrigal, a un costado de Ciudad Universitaria, así como las variables meteorológicas registradas en la estación del CCA de la Red Universitaria de Observatorios Atmosféricos (RUOA) .

Conocer las condiciones y calidad del aire que respiramos tiene gran relevancia en la salud pública, por lo que es importante generar y buscar herramientas que nos ayuden a la toma de decisiones sobre la regulación de fuentes de emisión. Dado que en la Ciudad de México se presentan frecuentemente altas concentraciones de PM_{10}, en este trabajo se pretende construir un modelo, implementado la estadística Bayesiana, alimentado por datos de las variables de

aforo vehicular y el uso de variables meteorológicas, con el fin de reproducir el comportamiento del PM con diámetro aerodinámico de 10 micras, PM_{10}. Este modelo podrá ser la base para realizarse en un futuro el pronóstico de dicho contaminante, con la finalidad de dar herramientas en el apoyo de las políticas públicas e informar sobre la calidad del aire a la población. Todo esto con el objetivo de disminuir la exposición de la población a material particulado.

En el Capítulo 1 de este trabajo se discute la problemática de la calidad del aire en las ciudades. Mientras que en Capítulo 2 se discuten trabajos previos sobre modelos de contaminantes atmosférico. Así mismo en este trabajo se describe y define el material particulado dentro del Marco Teórico, Capítulo 3, de igual modo se explica y describen las principales variables meteorológicas que interactúan directamente con la contaminación atmosférica. Al final del Capítulo 3 se discuten los principales enfoques de la estadística y se profundiza en el teorema de Bayes, siendo esto último el núcleo del trabajo. En el Capítulo 4 se mencionan los instrumentos y datos empleados para la elaboración del modelo. El Capítulo 5 se explica la elaboración del modelo con el enfoque Bayesiano. Por último, en el Capítulo 6 y 7 se muestra y discute el modelo multivariable de regresión lineal con enfoque Bayesiano.

1.1. Objetivos

El objetivo general de este trabajo es construir un modelo con base en la estadística Bayesiana, que pueda reproducir el comportamiento de las concentraciones de material particulado observadas en la estación del CCA para la elaboración, en un futuro, del pronóstico de dicho contaminante dentro de zonas urbanas . Los objetivos particulares del presente trabajo son:

I Encontrar correlaciones entre las concentraciones de PM_{10} y algunas variables meteorológicas en el sitio de estudio, y determinar cuales de éstas son más significativas.

II Identificar si existen correlaciones entre las actividades antropogénicas y la concentración de PM_{10} durante el día.

III Encontrar el proceso más eficiente para generar el modelo dentro del enfoque Bayesiano.

Capítulo 2

Antecedentes

En el informe anual del 2016 de calidad del aire en la Ciudad de México, que elabora la Red Autónoma de Monitoreo Atmosférico (RAMA), reporta el promedio anual de concentraciones de PM_{10} con 38 $\mu g/m^3$ (Autores, 2017), mismos que se encuentran justo en el límite de exposición que estableció la OMS (OMS, 2016), véase tabla 1.1. Dado que existe una estrecha relación entre la exposición a altas concentraciones de material particulado (PM_{10} y $PM_{2.5}$) y la ocurrencia de enfermedades respiratorias (OMS, 2016), es de vital importancia el desarrollo y el uso de modelos, para estimar las concentraciones de PM. Ya que esta información es de gran interés para perfeccionar el proceso de toma de decisiones y en los sistemas gubernamentales de gestión de la calidad del aire (Sfetsos et al., 2010) y dada la relevancia de este problema, hay un gran número de aplicaciones que permiten estimar las concentraciones de PM_{10} y $PM_{2.5}$. Sin embargo, el pronóstico de las concentraciones de dichas partículas es aún un problema abierto y se requiere de buenos modelos como base para aumentar la eficiencia del pronóstico.

A continuación se mencionan algunos trabajos que recrean el comportamiento de ciertos contaminantes atmosféricos usando diferente modelos matemáticos:

Kukkonen et al. (2003), construyeron escenarios futuros sobre las concentraciones de NO_2 y PM_{10} sobre la ciudad de Helsinki. Usando tres modelos: modelo de redes neuronales (NN por sus siglas en ingles), modelos estadístico lineal y sistema de modelado determinístico (DET) para "pronosticar"NO_2 y PM_{10}. Encontraron que el modelo NN mostraba un mejor desempeño, principalmente sobre las concentraciones de NO_2. Donde su media anual obtenido y sus desviaciones estándar varían de $0{,}86 \pm 0{,}02$ a $0{,}91 \pm 0{,}01$. Por lo que concluye que los NN pueden ser herramientas útiles para evaluar las concentraciones de NO_2 y PM_{10} en un intervalo

de tiempo sobre las áreas urbanas. Pero tiene como limitante, la distribución espacial.

Por otro lado, Hooyberghs et al. (2005) también emplean NN para pronosticar durante 24 horas promedios diarios de concentraciones de PM_{10} sobre Bélgica, comparan dos modelos; el primero es con concentraciones PM_{10} y altura de la capa límite (BLH, por sus siglas en inglés). En el segundo agregan, dirección del viento, nubes (cobertura) y día de la semana, lo que aumenta la precisión del modelo. Determinan que las fluctuaciones diarias en las concentraciones de PM_{10} sobre las zonas urbanas en Bélgica, se deben principalmente a las condiciones meteorológicas, y en menor medida a los cambios en las emisiones antropogénicas.

Por último, Sfetsos et al. (2010), usan redes neuronales con un Modelo Local con Algoritmo Cluster (LMCA) y un Algoritmo Cluster Híbrido (HCA, por sus siglas en inglés) para conocer las concentraciones de PM_{10} sobre Helsinki y Londres. También comparan una aproximación que ellos desarrollaron "localized line modelling"para el modelo de regresión lineal y una red neuronal artificial (ANN), estos autores no encuentran diferencia significativa entre los dos modelos. Pero implementando la aproximación que desarrollaron, obtuvieron resultados que disminuían el error de predicción significativamente, en comparación con los métodos convencionales en al menos un orden de magnitud.

Ya se mencionaron algunos de los trabajos donde se han elaborado modelos que se ajusten al comportamiento de concentraciones de PM. Sin embargo, en la literatura existen pocos trabajos, de modelos elaborados con estadistica Bayesiana, que describan el comportamiento de PM. Pero existen diversos trabajos donde aplican el enfoque Bayesiano para determinar diferentes caracteristicas de PM, tanto daños a la salud, como su distribucion espacial, entre otros: Le Tertre et al. (2005), Chen et al. (2018), Beloconi et al. (2018), Weber et al. (2016) son algunos trabajos donde aplican la estadística Bayesiana. Dandole enfasis al trabajo de Yu et al. (2015) en la que se basa parte de este trabajo.

Capítulo 3

Marco Teórico

3.1. Material Particulado

Los PM son de gran interés para la comunidad científica que se ha acrecentado gracias al avance tecnológico de los instrumentos de medición, con lo que se ha desarrollado un importante avance en el conocimiento de la composición química y las propiedades físicas del PM (Tomasi et al., 2017).

PM son emitidos a la atmósfera por fuentes naturales como el arrastre de arena y polen, por el viento, erupciones volcánicas, oleaje del mar, y por actividad antrópica como la quema de combustibles o biomasa. Ya sean emitidos directamente a la atmósfera (aerosoles primarios) o generados a partir de procesos de gas-partícula (aerosoles secundarios), los PM tienen un amplio rango de tamaños. Por lo tanto, se clasifican con su diámetro aerodinámico, que corresponde al diámetro de una esfera uniforme y esta determinada por la forma y densidad de la partícula. Los tamaños del diámetro aerodinámico va desde los nanómetros (*nm*) hasta unas decenas de micrómetros o micras (μm) (Tomasi et al., 2017). Así PM_{10}, $PM_{2.5}$ tienen diámetros aerodinámicos iguales o menores a 10 μm y 2.5 μm respectivamente. Ya que en este trabajo se utilizó concentraciones de PM_{10}, a partir de ahora usaremos por simplicidad PM para referirnos a PM_{10}.

Como se mencionó antes, existen diversas fuentes de aerosoles atmosféricos, así como diferente composición químicas o compuestos, los cuales pueden llegar a ser muy tóxicos para el ser humano. Por ejemplo, las partículas de plomo que son emitidos por las fábricas de pilas

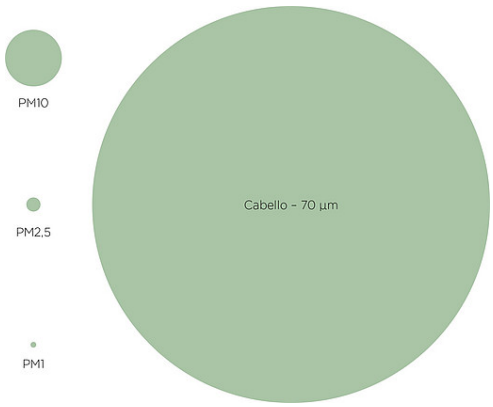


Figura 3.1: Comparación de tamaños de partículas (https://goo.gl/6VhE2H).

y pintura, procesos de construcción de edificio y de la quema de algunas gasolinas. Llegan a depositarse en la superficie rápidamente, contaminando suelo, comida, agua y llegan a ser ingerido por animales y humanos, acumulándose en el cuerpo teniendo efectos graves en la salud (Ahrens, 2012).

Las emisiones de los vehiculos se deben a la quema del combustible que requieren para funcionar y emiten diferentes contaminates a la atmósfera, entre ellos el material particulado. Las emisiones vehiculares dependen de las características del vehiculo, su tecnología, y su sistema de control de emisiones (de Ecología y Cambio Climático, 2009b). Los factores de emisión de PM para vehículos pesados de diesel son considerablemente mayores que los correspondientes a los vehículos y camiones ligeros que usan gasolina (de Ecología y Cambio Climático, 2009c). Sin embargo, en la Ciudad de México tiene mayor contribución de PM emitidos por vehículos a gasolina (de Ecología y Cambio Climático, 2009a), debido a la gran densidad de estos vehículos que transitan diariamente en la ciudad.

El material particulado es fácil de distinguir sobre las ciudades, debido a la reducción de la visibilidad que provoca por el exceso de partículas suspendidas en la atmósfera. $PM_{2.5}$ se toma como referencia de la calidad del aire en las zonas urbanas y es comúnmente monitoreado en todo el mundo, para analizar las relaciones en estudios epidemiológicos de la población vulnerable (Oprea et al., 2017). También se ha visto que el PM_{10} se presenta en mayores concentraciones en las temporadas invernales, debido a las altas emisiones de calentadores y quema de combustibles (Tomasi et al., 2017).

3.2. Ciencias de la Atmósfera

La atmósfera es una mezcla de gases que esta en continuo movimiento, este movimiento es producido por complejos flujos que llegan a abarcar desde algunos metros hasta cientos de kilómetros, dando lugar a fenómenos como la lluvia, nevadas, huracanes, entre otros que impactan directamente al ser humano en su vida diaria, por lo que se ha generado toda una ciencia que sea capaz de entender y explicar dichos fenómenos, a la que llamamos Ciencias de la Atmósfera, misma que nos ayuda a explicar el comportamiento de los contaminantes en la atmósfera.

Las Ciencias de la Atmósfera nos proporciona herramientas que son capaces de estudiar la emisión y formación de contaminantes secundarios, así como el tiempo que permanecen suspendidos los aerosoles en la atmósfera (Seinfeld and Pandis, 2016). Dado lo anterior podemos intuir una relación entre la meteorología y la calidad del aire.

A continuación se explican algunas variables meteorológicas y su influencia sobre la calidad del aire:

3.2.1. Capa Límite

La tropósfera es la capa de la atmósfera que se encuentra más cerca de la superficie de nuestro planeta, y es donde se presentan la mayoría de los fenómenos atmosféricos. Sin embargo, que esté en contacto con la superficie de la Tierra implica que hay una retro-alimentación entre los procesos y las características de la superficie(la evaporación , transpiración, emisión de contaminantes, etc) y los flujos atmosféricos. Un claro ejemplo de esto es el calor de la superficie: durante el día el Sol calienta con mayor eficiencia la superficie de la tierra que al aire, por lo que el aire se calienta por contacto con el suelo, así aumenta la temperatura de la parte más baja de la tropósfera, generando flujos de aire que dan origen a la advección y turbulencias, estos dos flujos son responsables del transporte de energía y materia en la atmósfera. A esta parte baja de la tropósfera se le llama **Capa Límite**. Por lo que la capa límite se define como la parte de la tropósfera que está directamente influenciada por la superficie terrestre y responde al forzamiento superficial al rededor de una hora o menos (Stull, 2012).

En la Figura 3.2 se observan las características principales que conforman la estructura de la capa límite a lo largo del día, donde se exhibe el desarrollo de los estratos en la capa límite

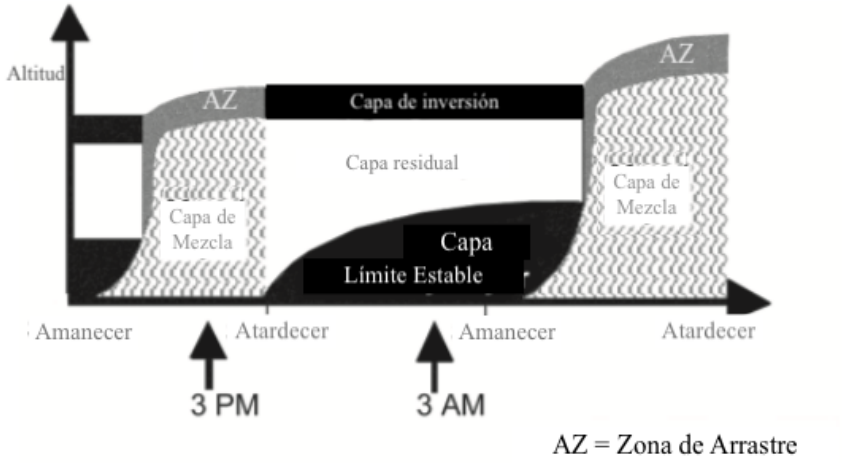


Figura 3.2: Estructura de la capa límite a lo largo del día (editado de Stull (2012).

en función de la altitud. En el día, cuando los rayos del sol calientan la superficie, esta capa aumenta su espesor, su altura, y también su movimiento turbulento, con ello existe una mayor dispersión de energía y contaminantes, al ocurrir esto, se le llama **Capa de Mezcla (CM)**. Sin embargo, cuando el sol se oculta se disipa gradualmente el calor y disminuye la turbulencia, quedando capas cuasi-estáticas o neutras, capas con residuos de turbulencia similares en la vertical y horizontal, a ésta se le llama **Capa Residual (CR)**. Cuando la atmósfera se enfría más, cerca de la medianoche, la turbulencia es suprimida y la dispersión es mayor en la horizontal que en la vertical, creando la llamada **Capa Estable (CE)**. Hay una capa estable en la parte superior de la CM que actúa como una tapa para las corrientes térmicas ascendentes, restringiendo así el dominio de la turbulencia, se le llama **Zona de Arrastre (ZA)**, porque el arrastre hacia CM ocurre ahí. A veces, esta capa estable es lo suficientemente fuerte y presenta una inversión de temperatura; es decir, la temperatura aumenta con la altura. Con frecuencia se le denomina **Capa de Inversión**.

La Capa Límite juega un rol determinante en la dispersión de los contaminantes. Casi todos los PM están contenidos en la Capa Límite, y están influenciados por la misma, por lo que el espesor de dicha capa determina el tamaño y concentración de los aerosoles (Wei-hua et al., 2017). Esto es fácil ver si tomamos la definición de concentración:

$$[x] = m_x/V, \tag{3.1}$$

donde $[x]$ es la concentración de la sustancia x, m es a masa de x y V el volumen de x. Así, entre más grande sea el volumen donde se encuentran los contaminantes, la concentración de

estos disminuye y si el volumen es pequeño la concentración aumenta.

3.2.2. Temperatura

La temperatura es una medida de la energía cinética de la materia. Sin embargo, la Temperatura del Aire en la Superficie (TAS) está determinada por el balance de la energía de la radiación solar que llega a la superficie y los flujos netos de calor sensible, latente y radiación de onda larga emitida por la superficie (Hughes et al., 2007). En otras palabras, la temperatura del aire está en función de la radiación del sol que calienta el suelo y la energía que absorbe y desprende el aire y suelo.

Desde que el sol sale por el horizonte llega radiación solar a la superficie terrestre, parte de esta energía es almacenada y otra tanta es emitida y reflejada a la atmósfera. Durante todo el día, el calor es continuamente almacenado por la superficie y aumentando la temperatura del aire, teniendo su máximo pasado el medio día, cuando el sol ha cruzado el cenit. Al atardecer, la radiación del sol disminuye, con ello la temperatura del aire en la superficie tiende a disminuir y llega a su mínimo justo antes del amanecer, volviendo a empezar el ciclo diurno.

Existe una variabilidad espacial en el ciclo diurno de la TAS, debida a las características del suelo, ya sea su humedad, color (reflectividad/absorción), cubierta vegetal y topografía . Esta variabilidad espacial del ciclo diurno de la TAS conduce a gradientes térmicos diurnos, esto genera gradientes de presión horizontal y ciclos diurnos en la circulación atmosférica (Hughes et al., 2007). Un claro ejemplo es la circulación de valle-montaña (o brisa costa/mar)

Por ello, la temperatura superficial del aire afecta directamente al espesor de la capa límite y a los gradientes horizontales, por lo tanto, la circulación atmosférica.

3.2.3. Humedad Relativa

La atmósfera es una mezcla de gases compuesta en su mayoría por nitrógeno (78 %) y oxigeno (21 %), el 1 % restante se compone de dióxido de carbono, vapor de agua y gases traza. A pesar de la escasa proporción de vapor en la atmósfera, en la troposfera juega un papel crucial en casi todos los fenómenos meteorológicos y climáticos de nuestro planeta. Es responsable de la precipitación de agua líquida, hielo, nieve y de mantener la temperatura estable, gracias a la absorción y emisión de radiación infrarroja(Ahrens, 2012).

Hay varias maneras de medir la humedad, pero instrumentalmente es más fácil medir la **Humedad Relativa (RH por sus siglas en inglés)**. La humedad relativa es la relación entre la presión parcial del vapor de agua e y la presión de saturación del vapor de agua $e_s(T)$ y se expresa en porcentaje:

$$RH = \frac{e}{e_s(T)} * 100, \tag{3.2}$$

las variables de presión que componen a la RH se obtienen a partir de la ecuación de Clausius-Clapeiron:

$$\frac{dP}{dT} = \frac{\delta S}{\delta V} = \frac{L}{T\delta V} = \frac{LP}{R_V T^2}, \tag{3.3}$$

donde S es la entropía, Pla presión, V volúmen. La ecuación 3.3 se puede aplicar a la presión de vapor de agua en presencia de aire (Andrews, 2010), por lo cual se puede escribir para la presión de vapor de agua como:

$$\frac{de_s}{dT} = \frac{Le_s}{R_V T^2}, \tag{3.4}$$

donde L es el calor latente de vaporización que en la atmósfera se puede aproximar a una constante. R_V es la constante de gas específica al vapor y T es la temperatura. Así mismo al integrar la ecuación 3.4 obtenemos la presión de saturación del vapor de agua:

$$e_s(T) = e_s(T_0) exp \frac{L}{R_V}(\frac{1}{T_0} - \frac{1}{T}). \tag{3.5}$$

El agua presente en el aire influye en el comportamiento de los contaminantes en la atmósfera. Algunas especies químicas que se encuentran en estado gaseoso llegan a solubilizarse con el agua. También, el agua en forma de precipitación deposita en la superficie a los aerosoles (deposición húmeda), así como ocurren reacciones químicas dentro de las gotitas de nubes y algunos gases y aerosoles, que a su vez sirven de núcleos de condensación que ayudan a la formación de nubes Seinfeld and Pandis (2016).

3.2.4. Radiación

Se llama luz visible a la luz que logramos percibir, que junto con las ondas de longitudes cortas (rayos gamma, X y ultravioleta) y longitudes de onda largas (infrarroja, microondas y radio) forman el espectro de radiación electromagnética. En la tabla 3.1 se muestran los intervalos de longitud de onda a las que pertenecen los componentes de la radiación electromagnética.

Radiación electromagnética	
Radiación	Intervalo de longitud de onda (m)
Gamma	$< 10x10^{-12}$
X	$200x10^{-9} - 10x10^{-9}$
Ultra violeta	$200x10^{-9} - 380x10^{-9}$
Visible	$780x10^{9} - 2{,}5x10^{-6}$
Infrarojo	$2{,}5x10^{-6} - 1x10^{-3}$
Microondas	$1x10^{-3} - 1$
Radio	> 1

Tabla 3.1: Tabla de intervalos del espectro electromagnético.

La transferencia de energía más importante que sucede en la atmósfera es debida a la radiación electromagnética (Liou, 2002), en su mayoría proviene del sol y una pequeña cantidad es emitida por la superficie terrestre (longitudes de onda en el infrarrojo). La radiación electromagnética viaja en forma de ondas y tiene una velocidad muy cercana a la de la luz en presencia de aire.

Las regiones del espectro asociados con la transferencia de energía en la atmósfera son la luz ultravioleta, visible, infrarroja y microondas (Liou, 2002).

El promedio de la energía neta absorbida por la atmósfera y la superficie es casi igual a la energía infrarroja radiada de regreso al espacio por el planeta. El calentamiento promedio anual debido a la radiación solar está influenciado fuertemente por la latitud, teniendo un máximo en el ecuador y mínimo en los polos. La radiación infrarroja saliente, por el contrario, casi no depende de la latitud; Por lo tanto, hay un excedente neto de radiación en la región ecuatorial y un déficit en la región polar. Este calentamiento diferencial de la atmósfera ecuatorial con relación a las latitudes más altas crea un gradiente de temperatura de polo al ecuador, produciendo inestabilidades baroclínicas que resultan con le transporte de calor hacia los polos (Holton and Hakim, 2012).

La presencia de los gases y partículas genera fenómenos físicos y químicos en la atmósfera por su interacción con la radiación solar, ya sea por dispersión o absorción de la radiación solar, se forman fenómenos ópticos como: arco iris, halos y auroras boreales. También, dada la interacción con los fotones y las moléculas de los gases, se producen reacciones fotoquímicas, donde se dice que existe fotolisis si la molécula se disocia. Este proceso juega un papel fundamental en la química de la atmósfera (Hobbs, 2000). Por la fotolisis se forma el ozono (O_3) en la tropósfera. El mecanismo de reacción del O_3 llega a ser muy complicado, pero se puede llegar a formar

a partir de la interacción con NO_2 y radiación solar con longitudes de onda en el intervalo de $0.4 - 0.625\mu m$ (Hobbs, 2000).

Otra función importante de la radiación y los constituyentes atmosféricos es el efecto invernadero, el cual es el calentamiento gradual de la atmósfera debido a gases y partículas que absorben la radiación infrarroja proveniente del sol y de la superficie de la Tierra.

3.2.5. Precipitación

El proceso de la formación de gotas de nube, puede esquematizarse de la siguiente manera, existen núcleos de condensación precursores de las gotitas y hielo, y la presencia de suficiente humedad en el ambiente para que se formen gotas de nube.

Las nubes están constituidas por pequeñas gotas de agua o cristales de hielo, para que estas se formen se requiere de sobresaturación del vapor de agua, por lo que es necesario la presencia de partículas suspendidas que activen la sobresaturación para la formación de gotas. A estas partículas suspendidas se les llama **núcleos de condensación (NC)**.

Las gotas de nube permanecen suspendidas por la corriente ascendente en la nube, debido a que su velocidad de caída en relación a la del aire es pequeña (Jonas, 1994). Sin embargo, cuando las gotas de nube crecen, pueden alcanzar tamaños suficientemente grandes para que las corrientes de aire ascendentes no logren mantenerlas suspendidas, formándose así la **precipitación** de agua líquida o hielo (nieve) (Jonas, 1994). Para que la precipitación de la nube llegue a la superficie, se necesitan condiciones de sobresaturación en el aire, así las gotas de lluvia o copos de nieve no se evaporen completamente antes de tocar tierra. Pero ¿cómo llegan a crecer las gotas e hielo de las nubes para que sean capaces de tocar la superficie terrestre? Existen dos principales mecanismos de crecimiento de gotas en la nube, que dependerá de si la nube esta compuesta de gotas de agua o partículas de hielo (Jonas, 1994).

Para explicar la formación de lluvia cálida (sin hielo) se requiere de condensación de vapor de agua a agua líquida, colisión de gotitas y coalescencia. La condensación requiere de un enfriamiento del ambiente, y el suceso más común que podemos pensar en la atmósfera es por expansión adiabática de una parcela de aire mientras va ascendiendo, ésta misma se expande hasta llegar a la altura donde se encuentra el punto de rocío, que es la temperatura y presión donde el vapor de agua hace la transición a agua líquida.

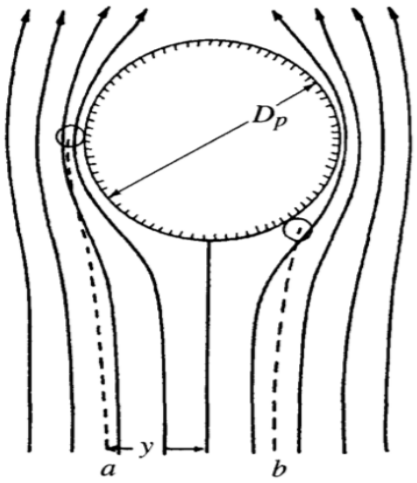


Figura 3.3: Esquema del flujo sobre una gota de nube cayendo. La línea punteada son las trayectorias de las gotitas más pequeñas (Seinfeld and Pandis, 2016).

La gotitas más grandes, con diámetro de D_p, caen más rápido que las de menor diámetro (d_p) (Seinfeld and Pandis, 2016), así mientras van cayendo, van impactando y atrapando a otras gotitas de menor tamaño, como se observa en la figura 3.3. Sin embargo, mientras cae la gota de D_p va empujando el aire, provocando que las fuerzas viscosas del flujo modifiquen la trayectoria (a y b) de las gotitas con d_p, alejándolas del centro de la gota grande como se observa en la imagen 3.3. Toda gota que se encuentre dentro del radio y de la gota D_p colisionará con ella, como el caso b, por el contrario, el caso con trayectoria a de la figura 3.3 únicamente roza a la gota grande.

La formación de lluvia fría, se refiere a la formación de partículas de hielo que conforman las nubes altas, un ejemplo son las cirrus. Estos cristales de hielo crecen por sublimación, deposición del vapor de agua, muy parecido al proceso de condensación en la lluvia cálida (Jonas, 1994).

La precipitación es una de las variables meteorológicas que interacciona más con los contaminantes, ya que al precipitar agua o hielo a la superficie, arrastra consigo los aerosoles que están suspendidos, removiéndolos de la atmósfera y depositándolos al suelo. También, las gotas de agua cambian su composición química debido a los gases y aerosoles presentes en la atmósfera (Hobbs, 2000).

3.2.6. Viento

El calentamiento que ocasiona la radiación solar (directa e indirectamente) impide que la atmósfera llegue al equilibrio, por lo que siempre se encuentra en continuo movimiento. Las diferencias de temperatura y presión crean inestabilidades en la atmósfera, generando flujos de aire tridimensionales, a este movimiento del aire es a lo que denominamos **viento**.

Conocer la naturaleza del viento ha sido difícil, sobre todo cerca de la superficie, donde su rugosidad, fricción y obstáculos provocan que el viento fluctúe de manera rápida y continua en la componente horizontal, a este movimiento se le conoce como turbulento. Por otro lado, el viento aumenta su velocidad con la altura, y la influencia de la superficie va disminuyendo mientras se incrementa la altitud. La velocidad promedio llega a ser constante entre los 500 y 2,000 metros de altitud respecto a la superficie (Strangeways, 2001). Así tenemos dos tipos de viento: turbulento y viento promedio. Los dos se encargan de transportar humedad, calor, momento y contaminantes a lo largo y ancho de la atmósfera, pero el viento promedio domina sobre la componente horizontal, mientras que el viento turbulento domina sobre la vertical (Stull, 2012). Ambos subgrupos pueden encontrarse dentro de la capa límite (Strangeways, 2001).

Las propiedades del viento se mide a partir de su dirección y su velocidad. La dirección del viento se mide por la dirección por donde sopla el viento, se expresa en grados en sentido a las manecillas del reloj tomando al norte como 00^{o} (Strangeways, 2001).

La velocidad y dirección del viento representan unas de las variables que más influyen en la concentración de los contaminantes, ya que se encarga de dispersarlos y está en función de su velocidad. También es fundamental el uso de la variable en los modelos de contaminantes, además de dispersar se puede saber las fuentes de procedencia de los contaminantes y tomar decisiones sobre las fuentes para evitar daños a la salud publica.

3.3. Enfoques estadísticos

Existen dos principales enfoques filosóficos en estadística. El primero es el llamado *enfoque Frecuentista*, también denominado *enfoque Clásico*. Dicho enfoque es el más conocido y usado, en todo libro de estadística básica se explica el enfoque Frecuentista. Se basa en la frecuencia del evento de interés y los procedimientos se desarrollan al observar cómo se desempeñan en todas las posibles muestras aleatorias. Las probabilidades no se relacionan con la muestra aleatoria obtenida (Bolstad and Curran, 2016). El segundo es el *enfoque Bayesiano* que parte del teorema de Bayes. Este aplicá directamente las leyes de la probabilidad al problema.

3.3.1. Enfoque Frecuentista

El enfoque Frecuentista tiene su origen en el siglo XIX, gracias a los trabajos de Karl Pearson y Francis Galton. El primero establece los principios de la estadística matemática, mientras que Galton descubrió el concepto de regresión a la media (Gutiérrez Peña, 2013). Posteriormente, a principios de siglo XX Ronal Fisher, Ergon Pearson y Lerzy Neyman plantean nuevos métodos para evaluar evidencias de observaciones al comparar dos hipótesis, hipótesis nula e hipótesis alternativa, basándose en la interpretación del enfoque frecuentista. Esta considera que la información provista por los datos D, es la única forma cuantificable de información probabilística, se usa como base para la construcción y evaluación de los procesos estadísticos bajo el comportamiento de frecuencias a largo plazo sobre repeticiones hipotéticas en circunstancias similares (Bernardo and Smith, 2001). El enfoque Frecuentista se basa, en general, en las siguientes ideas:

- Los **parámetros**, características numéricas de la población, son elementos constantes, fijos y desconocidos.
- Las probabilidades se interpretan como frecuencias relativas a largo plazo.
- El proceso estadístico se juzga a partir de la realización de un número infinito de repeticiones hipotéticas del experimento.

Al asignar los parámetros como fijos, obtienes una muestra aleatoria de la población para calcular una muestra estadística. Así, las probabilidades no se relacionan con la muestra aleatoria

particular, mismas que son adquiridas de la población (Bolstad and Curran, 2016). También el enfoque Frecuentista asocia a cada evento una probabilidad de obtener un valor verdadero del mismo, en base de frecuencias relativas. Por lo que no se le puede asignar una probabilidad a un evento que no ha ocurrido, por lo que también se le llama estadística objetiva.

Los métodos Frecuentistas requieren la decisión de rechazar o no una hipótesis, que simplemente es un reflejo del tamaño de la muestra. Si se tiene una muestra muy pequeña, lo más seguro es que no se pueda tener conclusión alguna. Si por el contrario, la muestra fuera muy grande, el rechazo a la hipótesis nula queda virtualmente asegurado (Silva and Benavides, 2001). Por otro lado, usando las pruebas de hipótesis se tiene que tomar una decisión entre dos opciones, por lo que no contribuye a valorar y proporcionar alguna credibilidad que estas puedan tener (Silva and Benavides, 2001).

3.3.2. Enfoque Bayesiano

Thomas Bayes murió en 1761 y se le recuerda por su trabajo *" An Essay Toward Solving a Problem in the Doctrine of Chances"* publicado de manera póstume por su amigo Richard Price en la *"Philosophical Transactions of the Royal Society"* en 1763. Este artículo sentó la idea de la estadística subjetiva mediante el teorema que lleva su nombre. Bayes mostró como la probabilidad inversa, puede ser usada para calcular la probabilidad de eventos previos a partir de la ocurrencia del evento consecuente (Bolstad and Curran, 2016). El teorema de Bayes fue utilizado por algunos científicos del siglo $XVIII$, entre ellos Laplace quien lo implementa de manera sistemática al análisis de datos (Malakoff, 1999). Las ideas en que se basa este enfoque son los siguientes:

- Los parámetros los consideramos aleatorios, ya que no estamos seguros o desconocemos el valor real de este.

- Las leyes de la probabilidad son usadas directamente para hacer inferencia sobre los parámetros.

- Las declaraciones que podamos hacer de la probabilidad sobre los parámetros deben ser interpretados como " grados de creencia ". La distribución *a priori* de probabilidad debe

ser subjetiva. Mide que tan plausible la persona considera que el valor de cada parámetro es antes de observar los datos(Bolstad and Curran, 2016).

- Se actualiza la información sobre los parámetros después de obtener resultados mediante el uso del teorema de Bayes (Bolstad and Curran, 2016). Como resultado, obtenemos una distribución *a posteriori* que nos da el peso relativo que le damos a cada parámetro, después de analizar los datos. Por lo tanto, esta distribución proviene de dos cosas: la distribución *a priori* y los datos observados (Bolstad and Curran, 2016).

De manera simplificada, el teorema de Bayes describe la probabilidad condicional para calcular la probabilidad de que ocurra un evento H a partir de observaciones D consecuentes a H. Si H la denotamos como un evento y D como observaciones previas, definimos el teorema como:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}, \tag{3.6}$$

donde consideramos $P(H)$ como una creencia de que ocurra H antes de tener información de los datos u observaciones D, del lado izquierdo de la igualdad lo consideramos como la creencia de H después de obtener los datos D. Por otro lado $P(D|H)$ y $P(D)$ nos plantea el mecanismo del teorema que proporciona una solución al problema sobre le aprendizaje de los datos. Como cualquier teorema de probabilidad, el teorema de Bayes proporciona una forma de " contabilizar"la incertidumbre (Bernardo and Smith, 2001).

Es en el siglo XIX, con el trabajo de Karl Pearson y su busqueda de la objetividad, se cuestiona un aspecto fundamental del trabajo de Bayes y Laplace, la subjetividad. La discusión y controversia que envuelve al teorema de Bayes es en principio a la interpretación filosófica de la probabilidad ¿objetiva o subjetiva? y su justificación se da con base a la teoría científica. Después, a principios y mediados del siglo XX, con los trabajos de Frank Ramsey, Bruno de Finetti y Leonard Savage le dieron una base firme a la teoría Bayesiana (Malakoff, 1999). Con esto se consolido el enfoque Bayesiano y se formalizo teóricamente la inferencia estadística a través de la teoría de la decisión (Gutiérrez Peña, 2013).

El uso del teorema de Bayes en la actualidad es bastante común y aceptado las pruebas de diagnóstico. En cambio, existe un activo debate en el uso del análisis estadístico general, donde los eventos de interés están en términos de los valores desconocidos de los parámetros de un

modelo y por lo que se requiere especificar una distribución de probabilidad sobre dichos valores (Gutiérrez Peña, 2013).

Es en este punto donde se ve la gran disparidad entre los métodos estadísticos clásico o frecuentistas y los Bayesianos. Mientras que la visión frecuentista se pregunta qué nos dicen lo datos acerca del valor del parámetro (ignorando la evidencia externa a los datos), definiendo la probabilidad en términos de experimentación, la frecuencia de eventos observados en relación a un número hipotético de experimentos. Los métodos Bayesianos se preguntan explícitamente cómo cambia nuestro estado de información acerca del valor del parámetro a la luz de los datos observados (Gutiérrez Peña, 2013), permitiendo asignarle una probabilidad *a priori*, $P(H)$, información basada en el conocimiento previo a la obtención de los datos, por lo tanto usa un grado de creencia previo de un evento dado.

El principal problema del enfoque Bayesiano, también por el cual es severamente criticado, es la subjetividad de elección de las condiciones iniciales, lo que resulta de conclusiones distintas a pesar de que se evalúen los mismos datos. Sin embargo, se tiene la ventaja que conforme se acumulen las evidencias u observaciones, la elecciones llegan a converger (Gutiérrez Peña, 2013). Mientras que el enfoque Frecuentista no cree que sea posible asignar un valor a un evento que no se conoce aún.

Los métodos Bayesianos reconocen que cada problema es distinto y promueven que el procedimiento de análisis se adapte al problema en cuestión (Gutiérrez Peña, 2013). No esta sujeto al tamaño muestral, si este es pequeño la información revelada sera pequeña. Entre más grande el tamaño de la muestra se podrá valorar mejor la información obtenida y mejor representará la realidad (Silva and Benavides, 2001). También tiene la ventaja que valorar la credibilidad por medio de la verisimilitud, con lo que evitamos las decisiones dicotómicas, como en el caso frecuentista, y con esto se puede actualizar la información previa a nuestro problema (Silva and Benavides, 2001). El uso de técnicas Bayesianas por lo general requiere de un esfuerzo computacional muy alto (Gutiérrez Peña, 2013). El cálculo preciso de la probabilidad *a posteriori* llega a ser complicado, ya que la mayor parte del cálculo se concentra en la distribución del parámetro de interés. Por lo que el uso de integrales es necesario para pasar de una distribución de probabilidad condicional a una colección de las distribuciones de probabilidad marginales (Bernardo and Smith, 2001), así lograr hacer inferencias sobre los parámetros (Gutiérrez Peña, 2013). En la mayoría de los casos, las integrales no pueden resolverse analíticamente, por lo que

es necesario el uso de métodos numéricos para aproximar dichas integrales(Gutiérrez Peña, 2013). El problema crece con el uso de grandes cantidades de parámetros. Los enfoques de producto cartesiano se vuelven computativamente imposibles y se requieren de enfoques alternativos para la integración numérica (Bernardo and Smith, 2001). Gracias a los avances tecnológico y teóricos a mitad del siglo *XX*, se desarrollaron técnicas numéricas flexibles y eficientes basadas en métodos de simulación estocástica (Gutiérrez Peña, 2013). A partir de esto se observa el florecimiento en el desarrollo y uso de los métodos Bayesianos.

3.4. Teorema de Bayes

El teorema de Bayes surge como consecuencia de relaciones matemáticas que obedecen las reglas o axiomas de probabilidad (Tabla 3.2), dicha relación parte del concepto de probabilidad condicional (ecuación 3.7).

Axiomas de la Probabilidad	
Axioma 1	$P(E) \geq 0$
Axioma 2	$P(\emptyset) = 0 \therefore P(\Omega) = 1$
Axioma 3	$SiE \cap F = 0, entoncesP(E \cup F) = P(E) + P(F)$

Tabla 3.2: Axiomas de la probabilidad.

Para cualquier evento G, dónde $G > 0$:

$$P(E|G) = \frac{P(E \cap G)}{P(G)}, \tag{3.7}$$

donde $P(E|G)$ es la probabilidad de que ocurra E dado el acontecimiento de un evento G, la ocurrencia del evento G arroja información sobre el evento E. $P(E \cap G)$ es la probabilidad conjunta de los eventos E y G y $P(G)$ es la probabilidad de G, sirviendo $P(G)$ como un factor de escala .

Se dice que un evento es independiente cuando su ocurrencia no afecta a otro evento; si un evento E es independiente de uno G, $E \perp G$, entonces la probabilidad conjunta o probabilidad de que ambos eventos ocurran es

$$P(E \cap G) = P(E)P(G). \tag{3.8}$$

Por lo tanto, si pasa que $E \perp G$ la ecuación 3.7 se vería de la forma:

$$P(E|G) = P(E) \leftrightarrow P(G|E) = P(G). \tag{3.9}$$

Si $E \perp G$ no fuera cierta, se dice que los eventos son dependientes.

Sabemos por teoría de conjuntos que $E \cap G = G \cap E$ ∴, así que la ecuación 3.7 la podemos escribir como:

$$P(E \cap G) = P(E|G)P(G) = P(E)P(G|E) = P(G \cap E). \tag{3.10}$$

Combinando las ecuaciones 3.9 y 3.10 obtenemos

$$P(E|G) = \frac{P(G|E)P(E)}{P(G)}. \tag{3.11}$$

Si $\{E_j \in \varepsilon, j \in J\}$ donde ε es una colección de eventos disjuntos, lo cual es la ocurrencia de uno de los eventos que excluye la ocurrencia del otro y viceversa (también llamados mutuamente excluyentes), tal que $P(E_j) > 0$. Y si $\{E_j \cap G, j \in J\}$, entonces

$$P(\bigcup E_j|G) = \sum_J P(E_j|G), \tag{3.12}$$

así, por la ecuación 3.10 expresamos 3.12 de manera más intuitiva y recordado que $G = \cup_j (G \cap E_j)$

$$P(G) = P(\bigcup E_j|G) = \sum_J P(G|E_j)P(E_j), \tag{3.13}$$

aplicando 3.13 y 3.10 en 3.7 obtenemos el teorema de Bayes

$$P(E_i|G) = \frac{P(G|E_i)P(E_i)}{\sum_J P(G|E_j)P(E_j)}, \tag{3.14}$$

que sí comparamos la ecuación 3.6 y 3.14, $E_j = H_j$, $G = D$. Donde por convención se les llama de la siguiente forma:

- $P(H_j), j \in J$, es la probabilidad a priori de $H_j, j \in J$:
- $P(D|H_j), j \in J$, es la verisimilitud de $H_j, j \in J$, dado G:
- $P(H_j|D), j \in J$, es la probabilidad a posteriori de $H_j, j \in J$:
- $P(D)$ es la probabilidad predictiva de D, son los datos o las observaciones.

Capítulo 4

Datos e Instrumentos

Se trabajó con datos meteorológicos, concentraciones de material particulado y de aforo vehicular, colectados cada minuto durante aproximadamente 8 horas (6:00 - 13:30), durante tres días (25, 26 y 27) del mes de marzo del año 2015. La base de datos meteorológicos y de material particulado de la estación de la Red Universitaria de Observatorios Atmosféricos (RUOA) y de la Red Automática de Monitoreo Atmosférico (RAMA), ubicados en la azotea del Centro de Ciencias de la Atmósfera (CCA) dentro de la Universidad Nacional Autónoma de México campus Ciudad Universitaria, al sur de la Ciudad de México. Los datos de aforo vehicular se tomaron a 300 m de distancia del CCA, sobre el puente de la estación del metro Universidad en donde registró el número y tipo de vehículos (autos particulares, transporte publico, taxis, etc.) que circulan sobre la avenida Delfín Madrigal, figura 4.1.

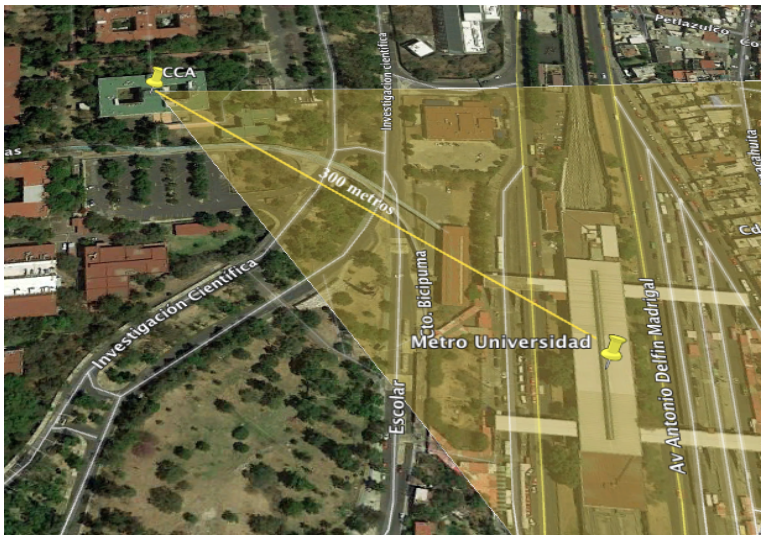


Figura 4.1: Sitios de recolección de datos meteorológicos, material particulado y aforo vehicular.

4.1. Instrumentación

A continuación se mencionan los diferentes instrumentos que se utilizaron para recolectar los datos meteorológicos, PM y aforo vehicular.

4.1.1. Instrumentos de medición meteorológicos

- **Sensor de temperatura y humedad exterior. Marca: Vaisala, Modelo: HMP155A** Este sensor monitorea la humedad relativa en un rango de 0 a 100 % RH y la temperatura en el rango de -80 a 60 ºC. Este sensor, esta compuesto por dos elementos, uno mide la temperatura y otro la humedad. La humedad relativa se mide con un sensor capacitivo, y la temperatura mediante un PRT (Platinum Resistance Thermometer). Figura 4.2a.

- **Sensor de radiación solar Marca Hukseflux, Modelo SR20-T1** Sensor basado en el uso de termopilas de alta precisión, mide la irradiación horizontal global (GHI) y tiene una respuesta espectralmente plana en todo el espectro solar. El SR20 genera una pequeña señal de voltaje que es proporcional al flujo de radiación solar, expresado en W/m^2. Figura 4.2b.

- **Anemómetro Marca: Gill, Modelo: Wind Sonic.** Wind Sonic es un anemómetro ultrasónico que mide la velocidad y dirección del viento. Se basa en el principio fundamental de la propagación del sonido respecto a la velocidad del viento, es decir, determina el tiempo que tarda en atravesar la señal de sonido en una distancia determinada, midiendo las fluctuaciones ocasionados por el viento. Figura 4.2c.

- **CLC Datalogger, Marca: Campbell Scientific Modelo CR1000.** El datalogger es el corazón de una estación de meteorológica, es el sistema de adquisición de datos y el control de los mismos. Mide las señales eléctricas de los sensores a una velocidad de muestreo establecida, procesa y almacena los datos. Figura 4.2d.

Instrumentos de medición de concentración de material particulado

Los datos de concentraciones de masa del material particulado (PM), se obtuvo de la base de datos de la Red Automática de Monitoreo Atmosférico, RAMA, localizada en el Centro de

(a) Sensor de temperatura y humedad relativa (imagen de https://goo.gl/6Nk8pV).

(b) Sensor de radiación solar SR20-T1(imagen de https://goo.gl/nQGpPR).

(c) Sensor de viento(imagen de https://bit.ly/2k2NKiw).

(d) Datalogger CR1000 (imagen de https://bit.ly/2L4EZ44).

Figura 4.2: Instrumentos meteorológicos.

Ciencias de la Atmósfera de la UNAM. El instrumento utilizado fue Thermo Fisher Scientific modelo FH62C14 el cual contabiliza las concentraciones de masa de PM_{10} y $PM_{2,5}$ en tiempo real.

El FH62C14 cuenta con un sensor de atenuación beta por el cual se puede obtener los promedios temporales de las concentraciones de masas de PM. Está diseñado con "dinamyc heating system"(DHS) para mantener la humedad relativa estable y que las partículas colectadas retengan el agua líquida y volátil que contengan. Figura 4.3.

4.1.2. Aforo vehicular

La base de datos del aforo vehicular, se generó a partir de contabilizar el tipo y cantidad de vehículos motorizados que pasaban en la avenida Delfín Madrigal, a un costado de la estación del metro Universidad, el conteo se realizó cada minuto en los días 25, 26, 27 de marzo del 2015

Figura 4.3: Thermo Fisher Scientific modelo FH62C14.

durante 8 horas cada día (6:00 a.m. a 1:30 p.m.). Esta base de datos fue proporcionada por el Dr. Arón Jazcilevich Diamant del Centro de Ciencias de la Atmósfera de la UNAM.

Capítulo 5

Metodología

5.1. Limpieza de los datos.

Con los datos obtenidos del aforo vehicular, meteorológicos y concentraciones de PM, se elaboró un programa en el lenguaje Python para efectuar la limpieza y tratamiento de los datos, del cual se obtuvieron un total de 19 variables entre meteorologicos y de aforo vehicular. Dichas variables se observan en la ecuación 5.4.

Ya que los datos de aforo vehicular se recolectaron el 25, 26 y 27 de marzo del 2015 a diferentes horarios cada día, se acoto cada variable al intervalo de tiempo del aforo, el cual resulto en los intervalos de 7:34 a 13:48 para el día 25, 6:55 a 14:19 para el día 26 y 7:07 a 13:20 para el día 27 de marzo.

Se agruparon los datos y se realizó el promedio horario de las variables meteorológicas y de concentración de PM. En el caso del viento, se hizo el promedio vectorial sobre las variables de dirección y velocidad del viento. En el caso del aforo vehicular, los grupos horarios obtenidos se realizó la suma por cada hora de cada una de las variables de aforo vehicular.

También se usó el método propuesto por Yu et al. (2015) para categorizar la variable de dirección del viento, asignando valores de 0 y 1 al viento a favor y viento en contra. En este caso se uso el 1 para asignar el viento con dirección a la estación meteorológica del CCA, que corresponde al area sombreada de la figura 4.1, con 88 a 157 grados donde el cero pertenece al norte.

5.2. Modelo.

Se utilizó una aproximación basada en la estadística Bayesiana para determinar los parámetros θ y realizar la selección del modelo. La ecuación de Bayes se ve de la siguiente forma:

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{l=1}^{k} P(D|M_l)P(M_l)}\ , \tag{5.1}$$

$$P(D|M_k) = \int P(D|\theta_k, M_k)P(\theta_k|M_k)d\theta_k$$

,

donde *D* son los datos;

θ_k es el vector de los parámetros en el modelo M_k;

$Pr(\theta_k|M_k)$ es la densidad $a\ priori$ de θ_k dentro de M_k;

$Pr(D|\theta_k, M_k)$ es la "likelihood"de los datos;

$Pr(M_k)$ es la probabilidad de priori de M_k; el mismo para todos los modelos.

Se considero un modelo de regresión lineal Bayesiano simple con distribución normal $a\ priori$ sobre los parámetros,

$$Y \sim N(\mu, \sigma^2), \tag{5.2}$$

Y es la varibale dependeinte, N la distribución normal cuyos estimadores son μ y σ^2, donde

$$\mu = \alpha + \sum \vec{\beta} X, \tag{5.3}$$

es la función lineal de la matriz de predictores **X**, α es la ordenada al origen y $\vec{\beta}$ el vector de los parámetros o coeficientes de covarianza de los predictores, mientras que σ representa el error de las observaciones. Así, en la ecuación 5.4 se muestra como se verian las variables, siendo PM la variable dependiente y el resto los predictores o variable independientes.

$$PM \sim \alpha + \beta_1 Motocicletas + \beta_2 Auto\_Particular + \beta_3 Camioneta + \beta_4 Taxi\_viejo + \beta_5 Taxi\_nuevo$$
$$+ \beta_6 Combi + \beta_7 Microbus + \beta_8 Autobus + \beta_9 Pick\_up + \beta_{10} Camión\_ligero + \beta_{11} Camión\_pesado$$
$$+ \beta_{12} Trailer + \beta_{13} Camión\_basura + \beta_{14} Temperatura$$
$$+ \beta_{15} HR + \beta_{16} Presión + \beta_{17} Radiación + \beta_{18} Categorias\_viento + \beta_{19} Velocidad\_viento + \sigma. \tag{5.4}$$

Se utilizo $a\ priori$ la distribución normal con media en cero y varianza 10 para ambos coeficientes de la regresión como aprecia en las ecuaciones 5.5 y 5.6, debido a que carecemos de información de ambos coeficientes les asignamos valores " pobres respecto a los valores reales que desconocemos de los parámetros. Ya que la varianza debe ser positiva, elegimos la distribución *Half-Normal* como el $a\ priori$ de σ , ecuación 5.7.

$$\alpha \sim N(0, 10), \tag{5.5}$$

$$\beta \sim N(0, 10), \tag{5.6}$$

$$\sigma \sim |N(0, 1), \tag{5.7}$$

Se utilizó la librería $PyMC3$, la cual está basada en el modulo $theano$ mediante el cual calcula gradientes a través de diferenciación automática. $PyMC3$ permite resolver problemas generales de inferencia estadística y de predicción bayesiana (Salvatier et al., 2016).

5.3. Ajuste del modelo y experimento (simulación).

En estadística Bayesiana, el maximum a posteriori (MAP) se usa estimar una cantidad desconocida con la misma dristribución a posteriori. Se empleó el MAP para optimizar la estimación puntual. La meta del MAP es encontrar θ que pueda maximizar la distribución de probabilidad del estimador, $P(\theta|y)$. La distinción es entre la θ según el cual los datos son más probables, y la mayor o mejor probabilidad de θ dados los datos (Ghoumari et al., 2018). Así mismo, se usaron dos métodos de optimización para encontrar los puntos mínimos de la función estimada dentro del MAP, estos fueron Powell y L-BFGS-B descritos por Buhmann (2017) y Liu and Nocedal (1989) respectivamente.

En el experimento se utilizo Markov Chain Monte Carlo (MCMC) para generar 5000 muestras del *a posteriori*, usando el algoritmo Metrópolis y No-U-Turn (NUT) incluida en la paquetería pymc3 de Python.

5.4. Highest Posterior Density

La selección de subconjuntos de variables predictorias es parte esencial de la elaboración del modelo de regresión lineal. El objetivo de la selección de variables se basa en lo siguiente: dado una variable independiente Y y un conjunto de posibles predictores $x_1, ..., x_j$, encontrar el mejor modelo de la forma:

$$Y = \alpha + \sum_{j=1}^{J} \beta_{ij} * x_{ij} + \varepsilon, \tag{5.8}$$

donde $x_{ij}, ..., x_{iJ}$ es un subconjunto de X y $\beta_{ij} \in \vec{\beta}$.

Para la selección de variables se uso el intervalo de *Highest posterior density (HPD)*. Construimos conjuntos de credibilidad dada la función de densidad $a\ posterirori$ del parámetro de la variable X_i, 5.9.

$$P(\vec{\beta_i} \in X_i|Y) = \int_X p(\vec{\beta_i}|Y)d\beta, \tag{5.9}$$

la probabilidad $100(1-\alpha)$ corresponde al intervalo simétrico del los percentiles $100(\frac{\alpha}{2})$ y $100(1-\frac{\alpha}{2})$ de la distribución $a\ posteriori$. El intervalo de $100(1-\alpha)$ % HPD es la región que satisface dos condiciones:

1. La probabilidad $a\ posteriori$ de esa región es $100(1-\alpha)$
2. La densidad mínima de cualquier punto dentro de esa región es igual o mayor que la densidad de cualquier punto fuera de esa región.

El HPD es el intervalo donde se encuentra la mayor parte de la distribución Wright (1986).

Capítulo 6

Resultados

6.1. Correlaciones

Se realizó un " heatmap" en donde se muestran las correlaciones de Pearson para todas las variables, se realizó una por cada día de datos observados y una donde se correlacionan todos los datos, figura 6.1. Las correlaciones del aforo vehicular son positivas respecto a las concentraciones de PM para los días 25 y 26 (figuras 6.1a y 6.1b), mientras que para el día 27 se observan en su mayoría correlaciones negativas respecto al material particulado (figura 6.1c). Por otro lado, la variable de meteorología que siempre tienen correlación positiva respecto al PM es la presión (figuras 6.1). Así mismo, la HR se observa que posee una correlación positiva en los días 25 y 26 pero el día 27 se ve una correlación negativa respecto al material particulado. También se realizó la correlación usando el enfoque Bayesiano, arrojando un intervalo del 95 % de HPD, mismo que es equivalente al obtenido en la figura 6.1 (ver Apéndice A.2).

Se sustituyó la variable de dirección del viento por categorías de viento, debido a que se quiere saber la procedencia del contaminante, si proviene de la avenida de Delfín Madrigal o no (0 y 1). En la figura 6.1a no muestra ninguna correlación para el día 25 entre categoría de viento y el PM. Por otro lado, en la figura 6.1b observamos una débil correlación para la misma relación de variables. Sin embargo, en la figura 6.1c tiene una correlación negativa respecto al PM, mismo día en que la mayoría del aforo vehicular tiene una correlación negativa. Lo anteriro sugiere que la contaminación provenga, en su mayoría, de la dirección contraría a Delfín Madrigal para el día 27 y de la avenida para los días 25 y 26. La categorización del viento en valores de 0 y 1 restringe a dos posibilidades (la contaminación viene de Delfín Madrigal o no),

(a) 25 de marzo del 2015.

(b) 26 de marzo del 2015.

(c) 27 marzo 2015.

(d) Correlación de los 3 días categorias.

Figura 6.1: Correlaciones entre todas las variables de los tres días.

razón por la que existirán días donde el viento sea a favor (igual a 0) por esto tendremos días con correlación nula como se observa en la figura 6.1a , o pueda presentarse únicamente en una hora del día y su correlación sea débil, figura 6.1b. Ya que la variable de categoría de viento nos brinda información útil, es necesaria incluirla al modelo.

6.2. Modelo

Para los parámetros α, β y σ usamos las distribuciones descritas en las ecuaciones 5.5, 5.6, 5.7, respectivamente como los $a\ priori$. La verosimilitud (likelihood) la asignamos como una distribución normal con μ siguiendo la forma de la ecuación 5.3.

Por otro lado, la estimación del MAP se realizó con dos métodos, Powell y L-FGSB-B, se calculó el coeficiente de determinación R^2 para estos métodos, siendo mejor el método de L-BFGS-B, tabla 6.1. Se realizó la simulación tomando en cuenta los valores obtenidos por el MAP y dos métodos numéricos diferentes: Metrópolis y NUT.

Se obtuvo un mejor coeficiente de determinación (R^2) al sustituir la variable de dirección por categorías de viento con ambos métodos, aumentando su R^2 en los dos, pero siendo mejor con el método de L-BFGS-B, tabla 6.1. Aunque ambos métodos se usan para la optimización de modelos multivariantes, pertenecen a diferentes métodos de optimización; directo en el caso de Powell e indirecto en el caso de L-BFGS-B. La principal diferencia entre uno y otro es que el primero utiliza valores sólo de la función objetivo en cualquier punto del espacio sin que sea necesario la diferenciabilidad de la función. Mientras que el segundo usa derivadas en la determinación de las direcciones de búsqueda reduciendo la función objetivo.

Hay dos características importantes en el método de L-BFGS-B que le dan ventaja sobre Powell: La primera es que L-BFGS-B ahorra el uso de memoria, generando computo rápido por su capacidad de de solo guardar y computar un número limitado de n-vectores (Huang et al., 2018), mientras que el método de Powell se limita a obtener dos direcciones conjugadas unidimensionales. La segunda característica es básicamente en la cantidad de variables que soporta L-BFGS-B (~ 500) (Gilbert and Lemaréchal, 1989).

Con los resultados mostrados en la tabla 6.1 se eligió el método de optimización L-BFGS-B. En la tabla 6.1 también se muestra el R^2 del modelo elegido usando la variable categórica del viento, se realizó con ambos métodos numéricos. En la tabla 6.1 vemos que tiene mejor ajuste

Método MAP	*Con vs Sin*	R^2		*MCMC*	R^2
Powell	Con categoría de viento	0.914		NUT	0.995
	Sin categoría de viento	0.86			
L-BFGS-B	Con categoría de viento	0.994		Metrópolis	0.991
	Sin categoría de viento	0.95			

Tabla 6.1: R^2 de los métodos y MCMC para el modelo.

NUT, pero este método tardo 6 veces($\sim$ 30 minutos) más que Metrópolis. Sin embargo, dado la distribución *a posteriori* de los parámetros (vease figura A.1) se determinó usar NUT.

El modelo consta de 5,000 simulaciones y las distribuciones del $a\ posteriori$; del lado izquierdo se muestra el la distribución que se obtienen en los parámetros dados los datos, también llamado Kernel Density Estimation (KDE-plot) de los parámetros, que es básicamente la versión continua de un histograma y en la parte derecha son todos los trazos (valores) que toman en función de los pasos muestreados, como se ven en la figura 6.2. Mientras que en la figura 6.3 se muestra la estadística descriptiva y el HPD de las distribuciones del a posteriori de α, β y σ obtenidos por el modelo. Los valores de las figuras 6.3c y 6.3d que se encuentran en cada nivel, corresponden a los β en cada uno de los predictores de la 5.4 respectivamente, por ejemplo: el valor de β promedio 0.572 de la figura 6.3c corresponde al parámetro que multiplica a la primera variable de 5.4, en este caso $Motocicletas$; el siguiente β 0.005 le corresponde a $AutoParticular$ y así sucesivamente. Lo mismo sucede con los HPD de la figura 6.3d.

En la figura 6.3d se muestras los parámetros de las variables del modelo, en el mismo orden que fueron incluidas dentro del modelo, teniendo como estructura el que aparece en la ecuación 5.4. Con los promedios de los parámetros arrojados por el modelo, figura 6.3d, y los valores horarios de las variables de 5.4, sustituimos los valores en la ecuación 5.8 de la regresión lineal multivariable y así ver la aproximación de la concentración de PM del modelo a los datos observados en el CCA, cuyo resultado se observa en la figura 6.4d. También se realizó el modelo con el promedio de cada 5 minutos de los datos, pero este mostró inconsistencias por la recolección de los datos (ver apéndice A.1), por lo que no fue implementado.

Las gráficas que se observan en la figura 6.4, en general, se aprecia muy buen ajuste entre la concentración de PM generados con el modelo versus las concentraciones de PM registras por la RAMA. Se llegan a ver ciertas diferencias entre estás gráficas. Para el día 27 se ve que el modelo casi empalma con los datos observados, figura 6.4c. Mientras que para el día 26 se

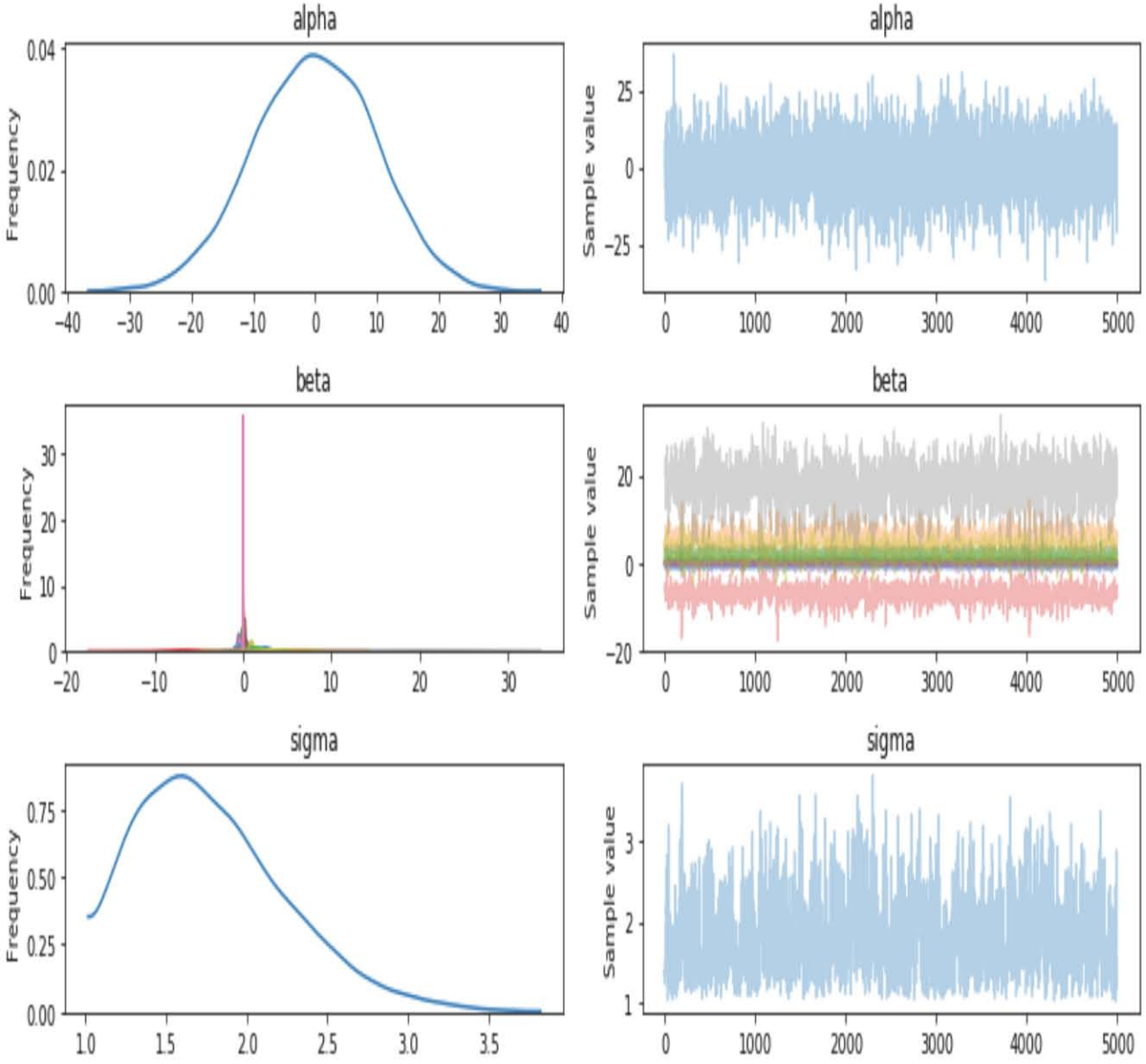


Figura 6.2: Distribución a posteriori de los parámetros del modelo.

alpha:

Mean	SD	MC Error	95% HPD interval
-0.186	9.812	0.141	[-19.006, 19.109]

Posterior quantiles:

2.5	25	50	75	97.5
-19.545	-6.866	-0.110	6.693	18.757

(a)

sigma:

Mean	SD	MC Error	95% HPD interval
1.791	0.467	0.019	[1.026, 2.662]

Posterior quantiles:

2.5	25	50	75	97.5
1.105	1.432	1.720	2.068	2.884

(b)

beta:

Mean	SD	MC Error	95% HPD interval
0.572	0.364	0.012	[-0.101, 1.324]
0.005	0.034	0.001	[-0.066, 0.072]
-0.129	0.120	0.004	[-0.387, 0.102]
0.046	0.072	0.003	[-0.092, 0.195]
-0.540	0.159	0.006	[-0.876, -0.238]
0.297	0.231	0.008	[-0.186, 0.735]
-0.396	0.220	0.006	[-0.821, 0.045]
-0.018	0.173	0.005	[-0.347, 0.331]
0.919	0.269	0.008	[0.362, 1.431]
-0.530	0.372	0.013	[-1.251, 0.205]
2.369	0.701	0.023	[0.971, 3.728]
5.838	1.803	0.060	[2.488, 9.599]
1.467	0.751	0.014	[-0.037, 3.026]
-7.013	1.896	0.062	[-10.833, -3.272]
-0.617	0.543	0.019	[-1.758, 0.383]
0.204	0.087	0.003	[0.028, 0.372]
-0.023	0.012	0.000	[-0.046, -0.001]
17.664	4.709	0.152	[8.693, 27.205]
2.662	1.609	0.037	[-0.819, 5.643]

(c)

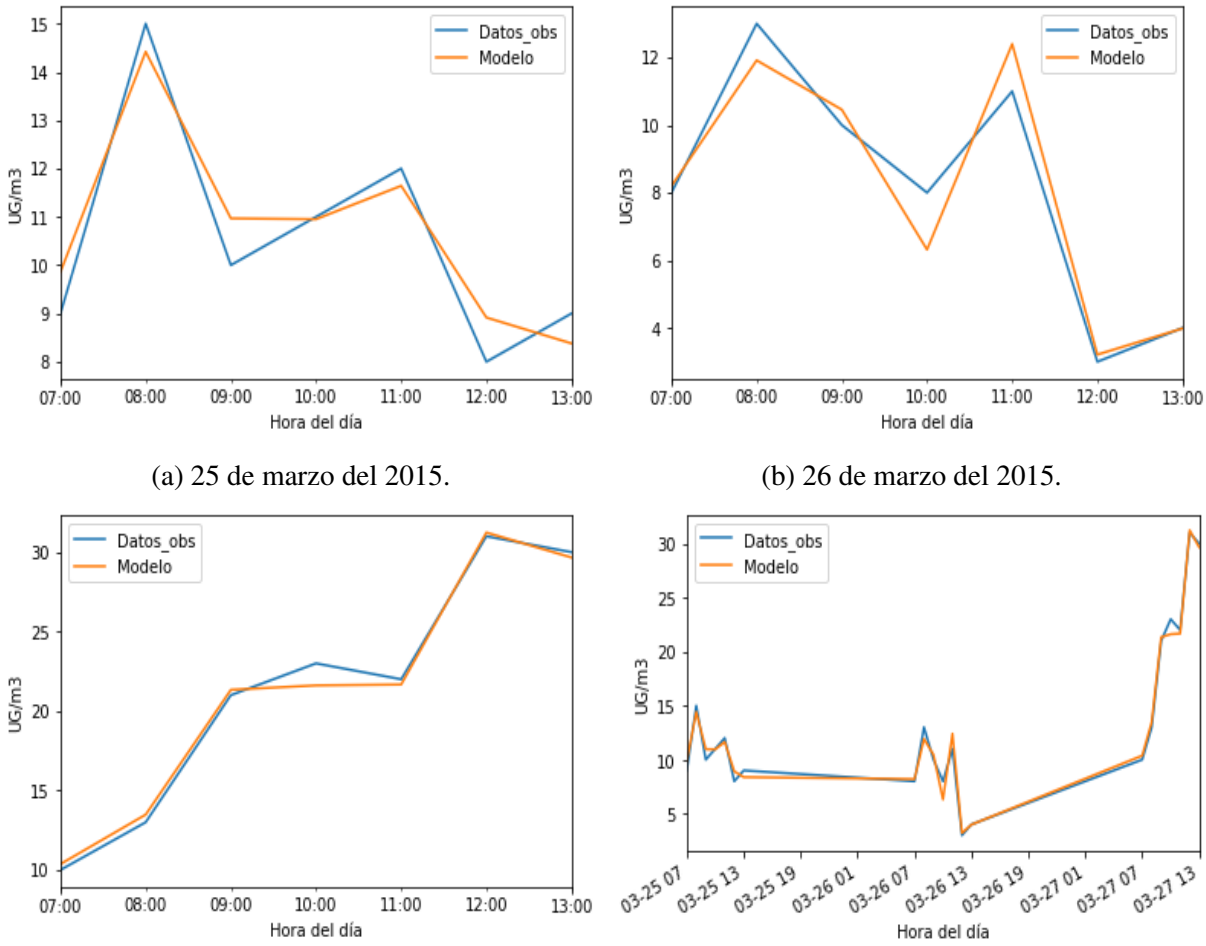
Posterior quantiles:

2.5	25	50	75	97.5
-0.087	0.332	0.553	0.779	1.373
-0.061	-0.018	0.004	0.026	0.078
-0.372	-0.202	-0.127	-0.057	0.119
-0.086	-0.000	0.041	0.087	0.203
-0.891	-0.633	-0.531	-0.436	-0.248
-0.210	0.161	0.311	0.452	0.716
-0.832	-0.528	-0.394	-0.261	0.040
-0.337	-0.134	-0.024	0.090	0.346
0.363	0.753	0.927	1.089	1.433
-1.211	-0.790	-0.548	-0.298	0.265
0.827	1.952	2.403	2.845	3.641
2.529	4.629	5.773	6.874	9.676
-0.087	1.020	1.476	1.932	2.988
-11.006	-8.136	-6.931	-5.821	-3.340
-1.804	-0.919	-0.581	-0.255	0.351
0.045	0.148	0.201	0.253	0.395
-0.048	-0.030	-0.022	-0.015	-0.002
7.448	14.705	17.956	20.895	26.349
-0.735	1.716	2.725	3.673	5.746

(d)

Figura 6.3: HPD y estadística descriptiva del a posteriori del modelo.

aprecian picos con mayor desface respecto a los datos observados 6.4b. Y para el día 25 la curva del modelo nunca rebasa los máximos de las concentraciones de PM observadas. Cabe recordar que el día 27 es el que presenta mayores correlaciones negativas entre el aforo vehicular y las concentraciones de PM, así como un valor igual a 1 en la categoría de viento. Mientras que el día 26 se tuvo una correlación débil entre la categoría de viento y una ligera disminución de las correlaciones de aforo vehicular. Por otro lado, el día 25 no influye la categoría de viento ya que esta tiene una correlación igual a cero. Lo anterior sugiere que el modelo tendrá un " peor .ªjuste si la correlación de categoría de viento tiene una correlación débil respecto al PM. Mejorando el ajuste cuando vale cero o siendo 1 categoría de viento. También el modelo nos dice que el comportamiento de las concentraciones de PM es bimodal para cada día de los tres estudiados.



(a) 25 de marzo del 2015.

(b) 26 de marzo del 2015.

(c) 27 marzo 2015.

(d) Todos los días estudiados.

Figura 6.4: Datos observados de concentración de PM vs datos del modelo para cada día.

Capítulo 7

Discusión y Conclusiones

El método propuesto por Yu et al. (2015) proporciona mayor información, al agregar la categoría de viento proporciona información de la dirección del cual proviene el contaminante, mejorando el ajuste del modelo. Así, conociendo las correlaciones entre concentraciones de PM, categoría de viento y el modelo de la figura 6.4c, nos sugiere que el aumento de las concentraciones provienen de la dirección opuesta a la avenida Delfín Madrigal. Así mismo, para el día 25 la correlación nos sugiere que las concentraciones de PM registradas, provienen de la dirección de dicha avenida. Mientras que para el día 26, el comportamiento de la concentración de PM influye de las dos direcciones, pero mayoritariamente de Delfín Madrigal.

De la distribución de los *a posteriori* de los parámetros y su intervalo del 95 % HPD, figura 6.3d, se puede interpretar el grado de credibilidad de las variables. Los parámetros que tengan promedios cercanos a cero dentro del 95 % HPD, sugieren que la variable influye muy poco en el comportamiento de las concentraciones de PM para los tres días, tal como: auto particular, autobús y taxi viejo. De estas variables, auto particular y taxi viejo tienen la mayor cantidad de vehículos totales (ver tabla A.1). Sin embargo, podríamos interpretar que el modelo sugiere que no son significativas. Esto es causa de la distribución de estas variables, al ser los vehículos de mayor uso en la Ciudad de México, no van a variar abruptamente a lo largo del día. Pero en el caso de taxi viejo, el intervalo del 95 % de HPD se encuentra entre [-0.092 , 0.195], donde la mayoria de los datos tienen valor positivo. Recordando que es la variable con mayor cantidad de vehículos (ver tabla A.1), por lo que este tendrá un valor del parámetro menor respecto al resto de variables. Por lo que es engañoso sólo fijarnos en su media y podríamos decir que influye

poco, mas no muy poco, en el PM. Así mismo, una de las variables meteorológicas que más influyen en el modelo, son la de la temperatura, presión y categoría de viento. Con signos muy parecidos a las correlaciones de la 6.1.

El modelo tiene un mejor ajuste en el día 27 de marzo que en el resto de los días (figura 6.4c), y parece tener relación a un mejor ajuste del modelo cuando se tienen concentraciones altas del contaminante. Lo anterior sugiere que el modelo tendrá un " peor .ªjuste si la correlación de categoría de viento tiene una correlación débil respecto al PM. Mejorando el ajuste cuando vale cero o siendo 1 categoría de viento. También el modelo nos dice que el comportamiento de las concentraciones de PM es bimodal para cada día de los tres estudiados.

Sin embargo, los métodos numéricos que se utilizaron tuvieron un comportamiento diferente en la densidad *a posteriori* de los parámetros, siendo NUT el que tuvo mejor ajuste, pero costando mayor tiempo en las simulaciones, $\sim$ 30 minutos más que Metrópolis. A su vez, Metrópolis no reproducía tan bien las distribuciones $a\ posteriori$ de los parámetros, como se observa en la figura A.1 (ver Apéndice A.2). Pero no hay mucha diferencia entre sus R^2, con lo que podríamos usar Metrópolis para disminuir el costo computacional y realizar, a futuro, un pronóstico de PM.

7.1. Conclusiones.

Se determino un modelo que reproduce bastante bien el comportamiento del contaminante estudiado con $R^2 = 0{,}995$, usando la regresión lineal multivariable con el enfoque Bayesiano, misma que nos ofrece cierta flexibilidad en el uso de variables independientes. Aunque el costo computacional pueda incrementarse dependiendo de la cantidad de parámetros que se utiliza, este problema disminuye usando métodos como L-BFGS-B y MCMC como Metrópolis.

También encontramos que las variables meteorológicas con mayor importancia son la temperatura, presión y categoría de viendo. Esta ultima nos proporciona la dirección de donde proviene el contaminante y mejora el ajuste del modelo. Por otro lado, la variable de auto particular tiene poca significancia en el modelo, a pasar de que representa la segunda varible con mayor cantidad dentro de los tres días.

En general el modelo tuvo un buen comportamiento y se encuentra dentro de la escala en que se obtuvieron las concentraciones del material particulado observado, gracias al enfoque Bayesiano y su ductilidad al tener poca disponibilidad de los datos. Así el tener pocas mediciones

del aforo vehicular y tener únicamente 21 datos generados en tres días, el modelo representó bastante bien los valores observados del material particulado. Se espera en un futuro realizar el pronóstico del contaminante con base a este modelo.

Apéndice A

Apéndice

A.1. Cantidad de vehículos

La tabla A.1 se muestra la cantidad del aforo vehicular que circulaban los días 25, 26 y 27 de marzo del 2015.

Cantidad de vehículos por día

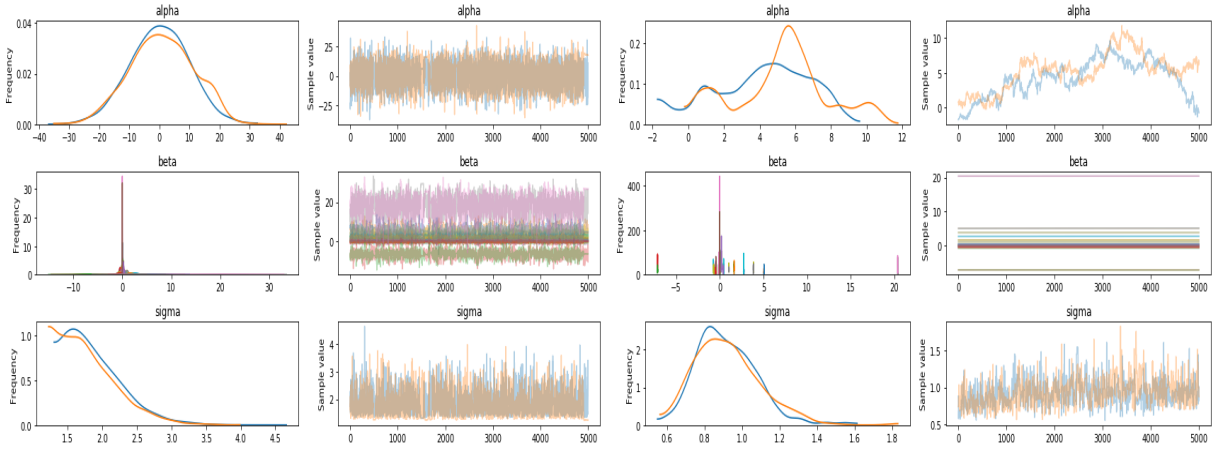
Aforo vehicular	**25 de marzo 2015**	**26 de marzo 2015**	**27 de marzo 2015**	**Total**
Motocicletas	152	164	184	500
Auto Particular	2031	2623	2380	7035
Camioneta	561	730	658	1949
Taxi viejo	2278	2720	2533	7531
Taxi nuevo	648	760	719	2127
Combi	645	852	761	2258
Microbús	330	416	406	1152
Autobús	522	634	590	1746
Pick up	145	154	152	451
Camión ligero	114	115	134	363
Camión pesado	43	46	46	135
Trailer	3	1	3	7
Camión de basura	7	7	7	21

Tabla A.1: Cantidad de vehículos por cada día y de todo el aforo vehicular.

A.2. Nut vs Metrópolis

La figura A.1 muestra las distribuciones $a\ posteriori$ de los parámetros generados por ambos métodos, NUT (figura A.1a) y Metrópolis (figura A.1b). Claramente se aprecia que el MCMC

NUT mejora el ajuste del modelo. Sin embargo, llega a tardar $\sim$ 20 minutos más que MCMC Metrópolis.



(a) Parámetros usando NUT .

(b) Parámetros usando Metrópolis.

Figura A.1: Distribución de los parámetros de los modelos generados.

A.3. Modelo con datos promediados cada 5 minutos.

También se realizo el modelo usando datos promediados cada 5 minutos. A pesar de tener datos cada minuto de todas las variables, los datos de PM no cambiaban en lapsos de tiempo prolongados, creando valores extremos en el modelo como se muestra en la figura A.3d, así como el ajuste del modelo sobre los parámetros están desfasados, figura A.2. El R^2 que se obtuvo en este modelo fue de $0{,}557$.

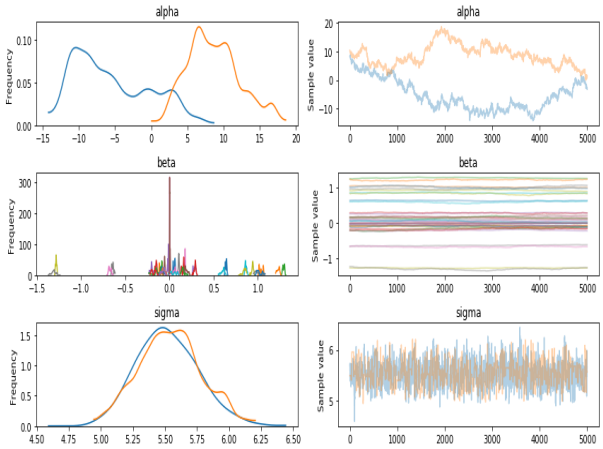
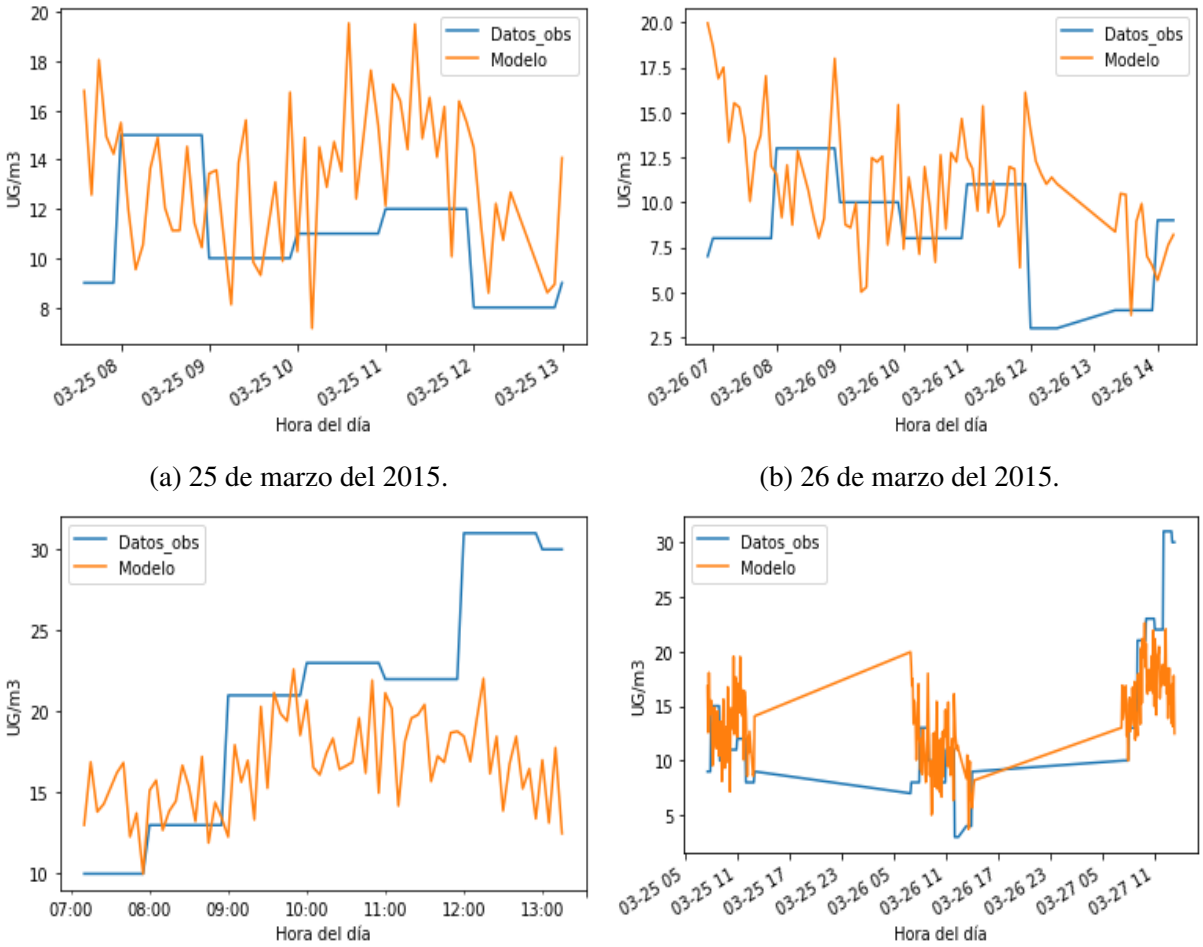


Figura A.2: Parámetros usando Metrópolis con datos cada 5 min .



(a) 25 de marzo del 2015.

(b) 26 de marzo del 2015.

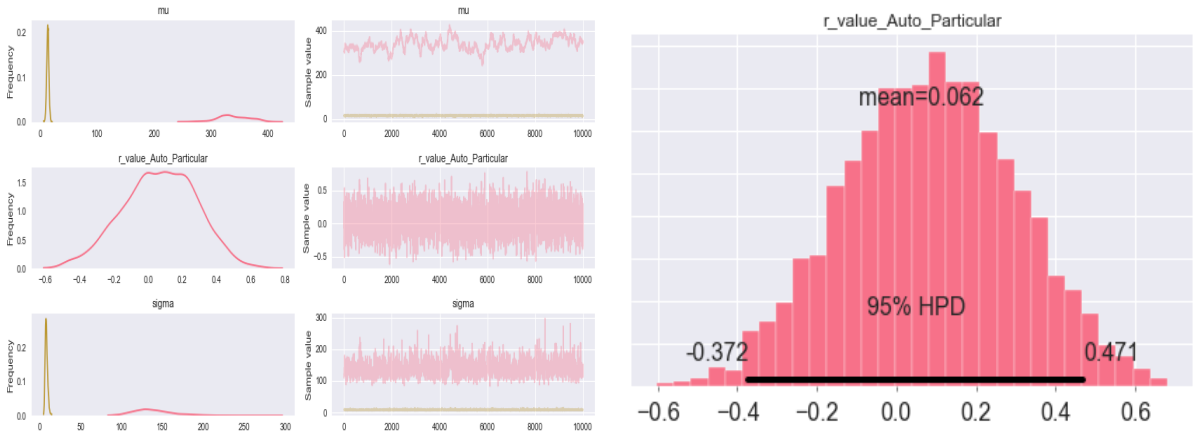
(c) 27 marzo 2015.

(d) Todos los días estudiados.

Figura A.3: Datos observados de concentración de PM vs datos del modelo para cada día.

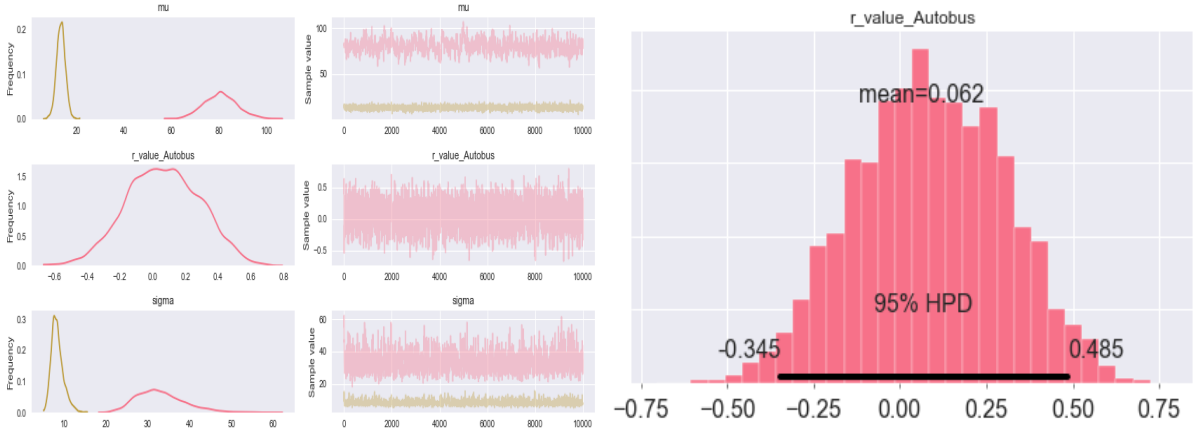
A.4. Correlaciones con Bayes.

Se realizaron los intervalos de HPD usando el método Bayesiano, entre las variables predicotrias y las concentraciones de PM, estas gráficas son análogas a la columna de $Material\ Particulado$ de la figura 6.1. La gran diferencia es que en la figura 6.1 asigna un único valor, mientras que en el enfoque Bayesiano es una distribucion con un intervalo de significancia como ya se ha discutido. A continuación se muestras las los $a\ posteriori$ (inciso a) y la distribución del 95 % HPD (inciso b) de cada una de las variables con la concentración de PM:



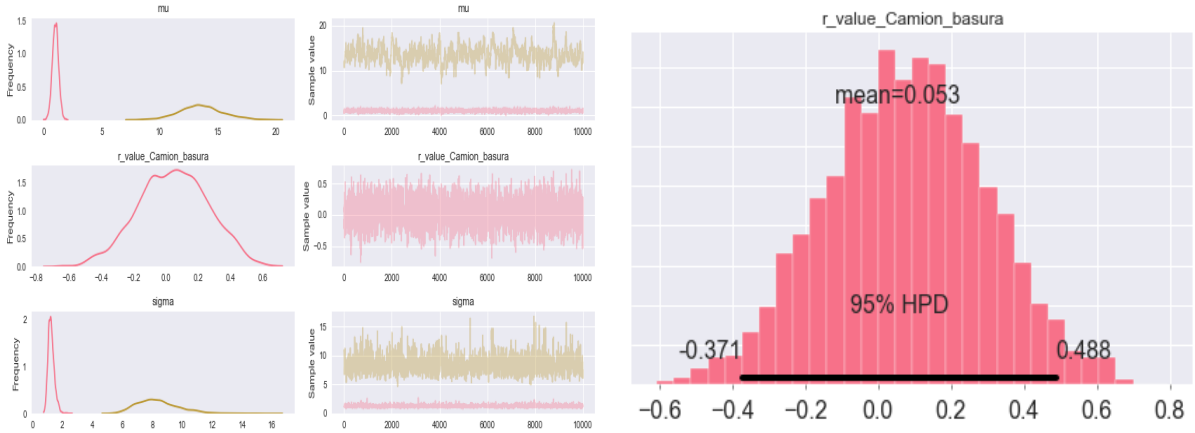
(a) *A posterioris* de los parámetros. (b) 95 % de HPD.

Figura A.4: Datos de Auto Particular.



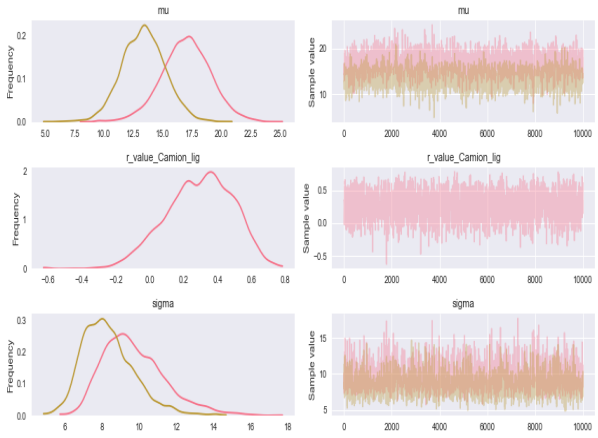
(a) *A posterioris* de los parámetros. (b) 95 % de HPD.

Figura A.5: Datos de Autobús.

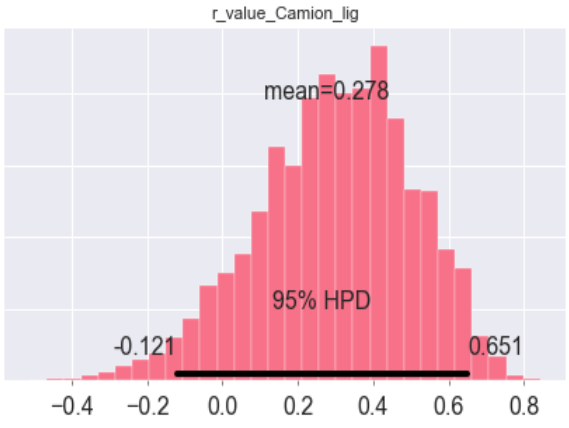


(a) *A posterioris* de los parámetros. (b) 95 % de HPD.

Figura A.6: Datos de Camión de basura.

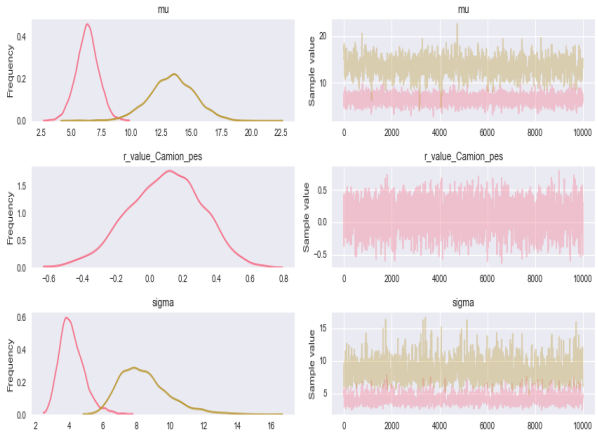


(a) *A posterioris* de los parámetros.

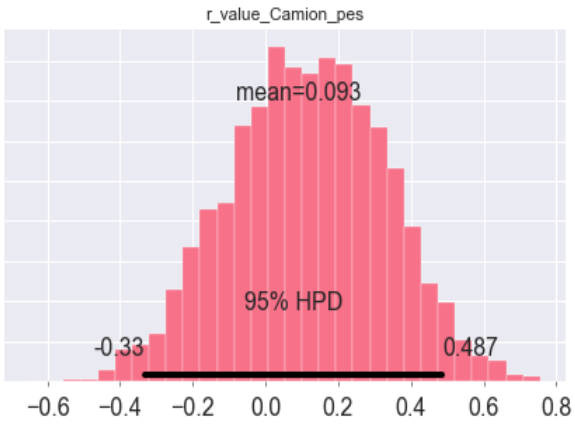


(b) 95 % de HPD.

Figura A.7: Datos de Camión Ligero.

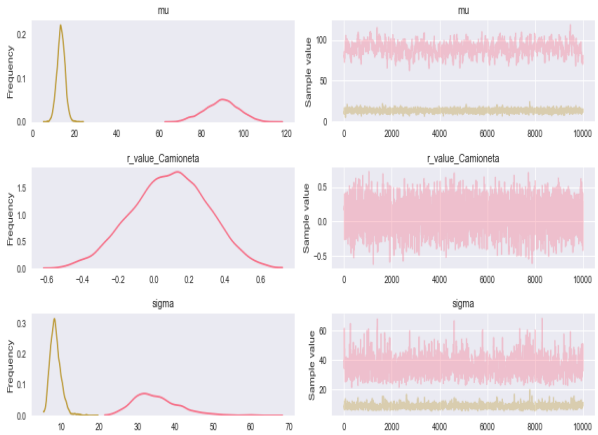


(a) *A posterioris* de los parámetros.

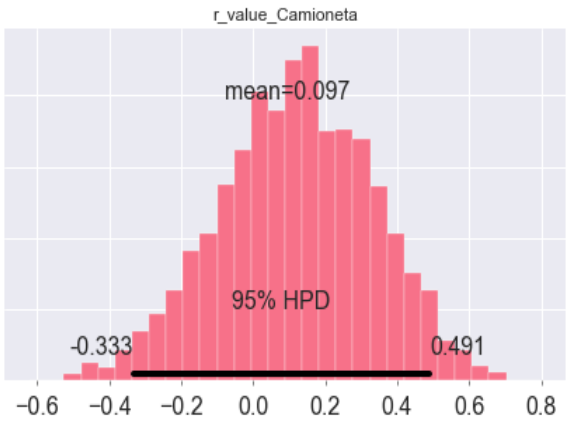


(b) 95 % de HPD.

Figura A.8: Datos de Camión Pesado.

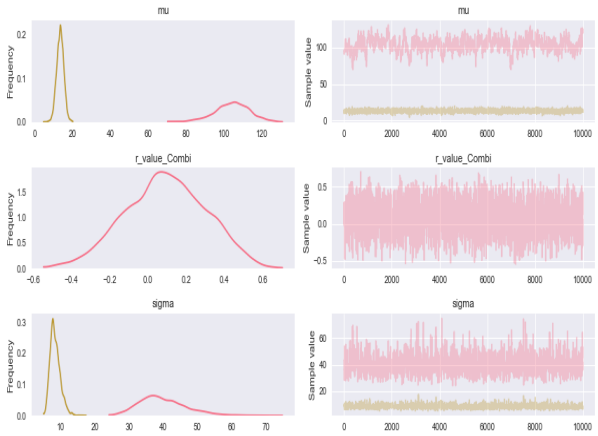


(a) *A posterioris* de los parámetros.

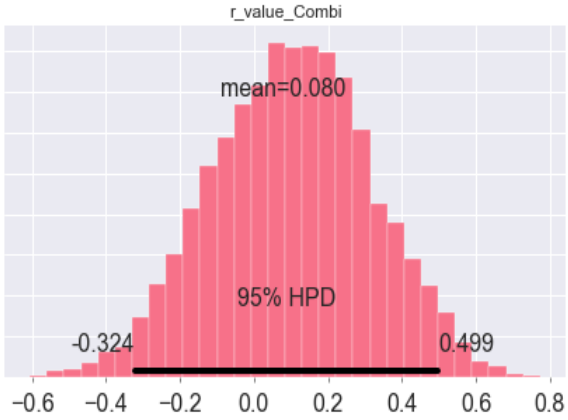


(b) 95 % de HPD.

Figura A.9: Datos de Camioneta.

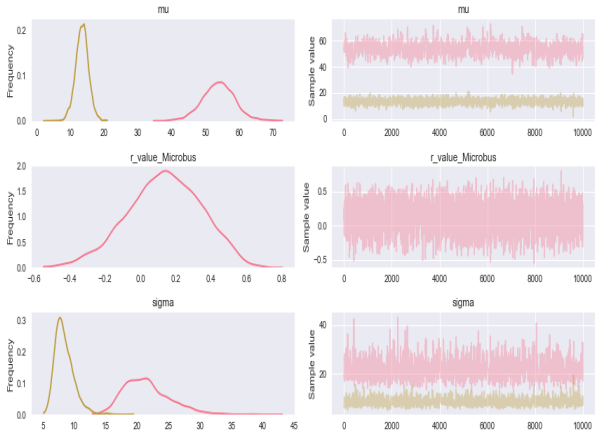


(a) *A posterioris* de los parámetros.

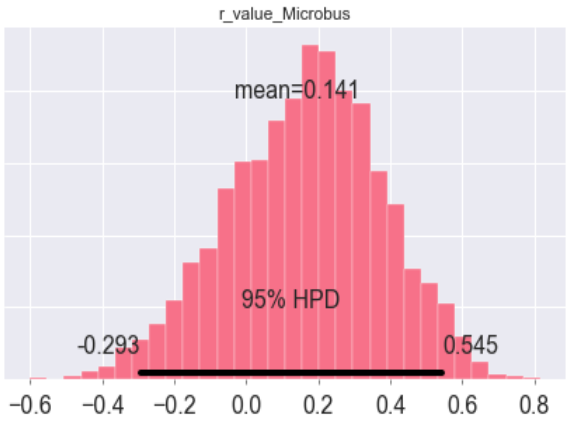


(b) 95 % de HPD.

Figura A.10: Datos de Combi.

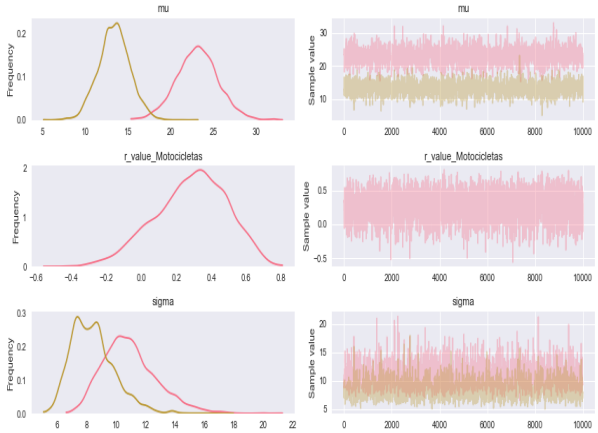


(a) *A posterioris* de los parámetros.

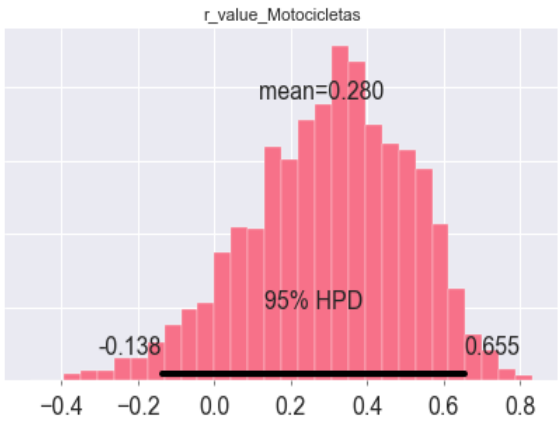


(b) 95 % de HPD.

Figura A.11: Datos de Microbus.

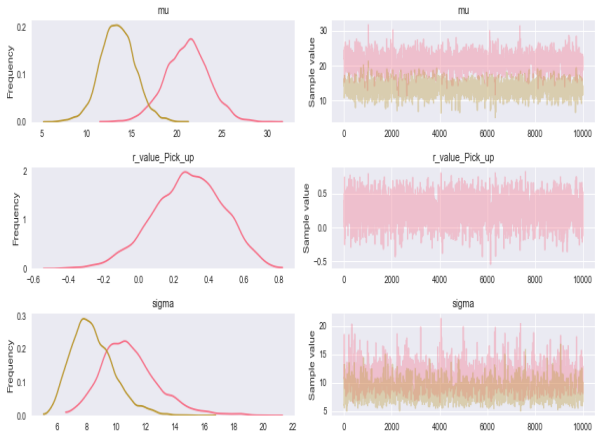


(a) *A posterioris* de los parámetros.

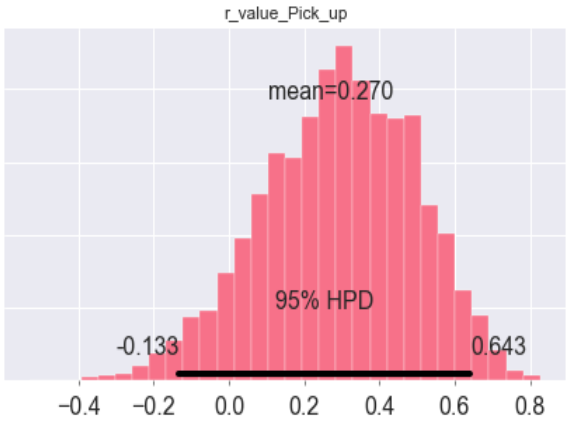


(b) 95 % de HPD.

Figura A.12: Datos de Motocicletas.

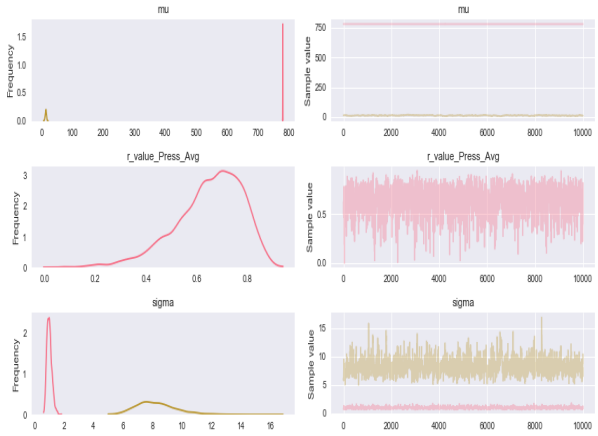


(a) *A posterioris* de los parámetros.

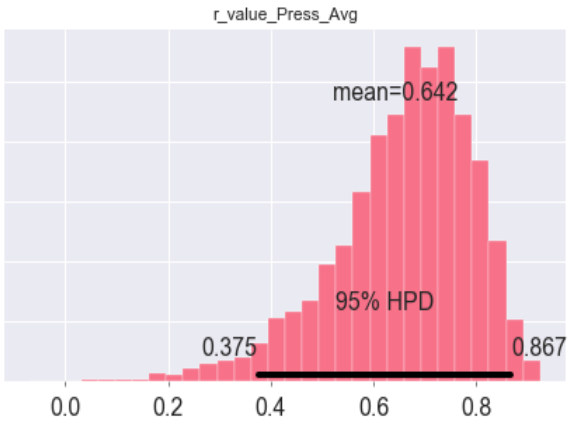


(b) 95 % de HPD.

Figura A.13: Datos de Pick up.

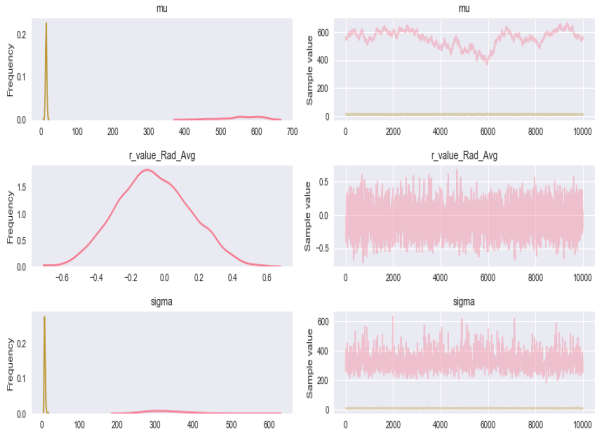


(a) *A posterioris* de los parámetros.

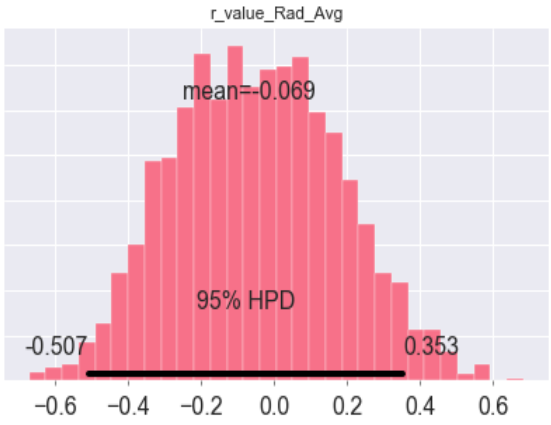


(b) 95 % de HPD.

Figura A.14: Datos de Presión promedio.

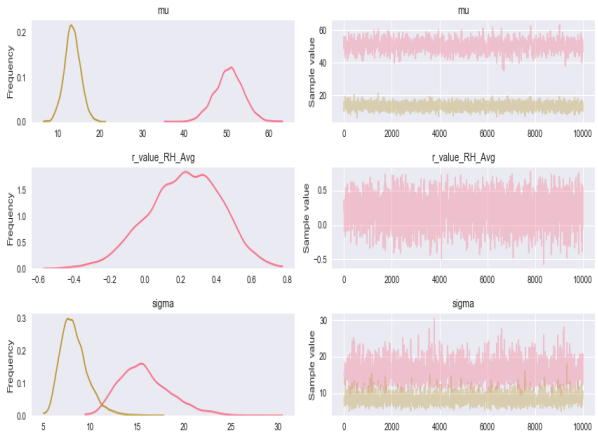


(a) *A posterioris* de los parámetros.

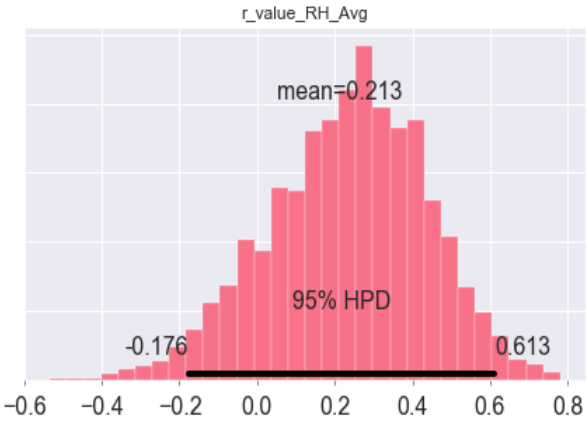


(b) 95 % de HPD.

Figura A.15: Datos de Radiación.

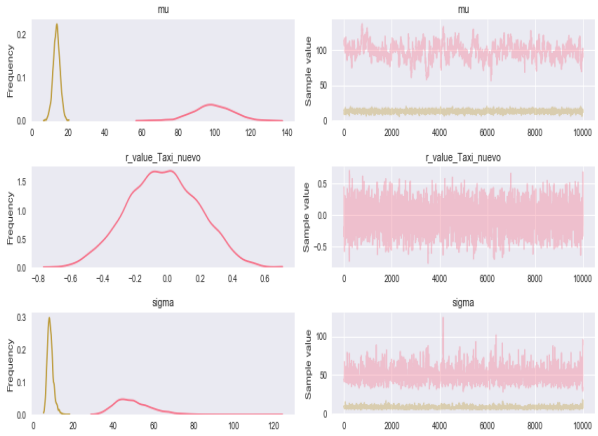


(a) *A posterioris* de los parámetros.

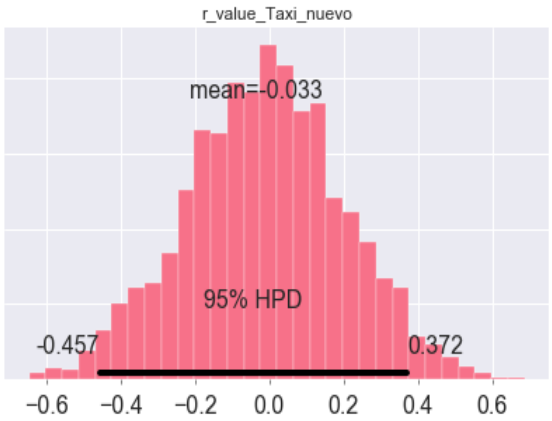


(b) 95 % de HPD.

Figura A.16: Datos de Humedad Relativa.

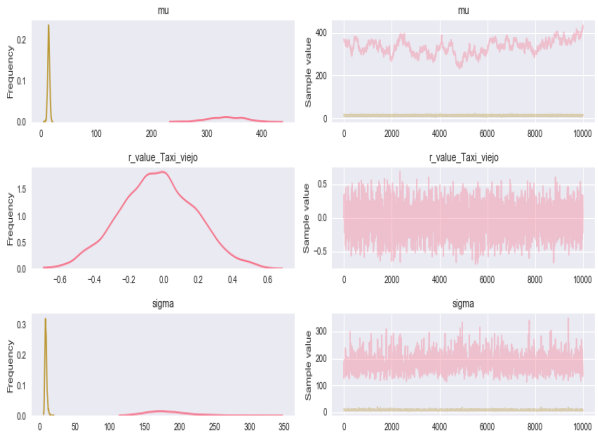


(a) *A posterioris* de los parámetros.

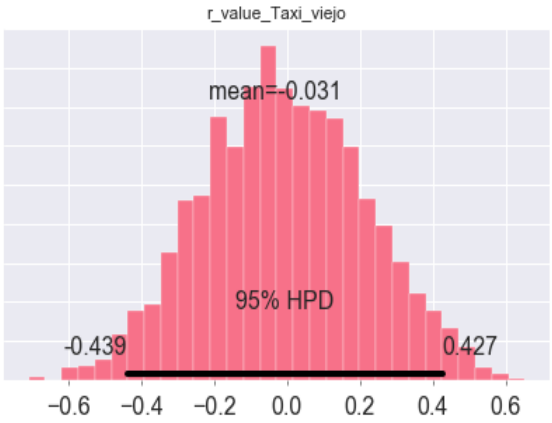


(b) 95 % de HPD.

Figura A.17: Datos de Taxi nuevo.

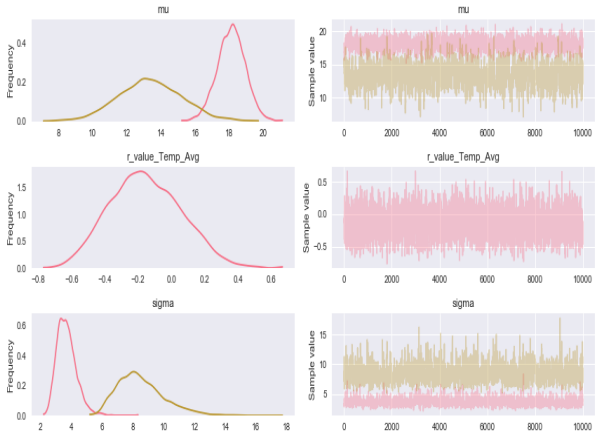


(a) *A posterioris* de los parámetros.

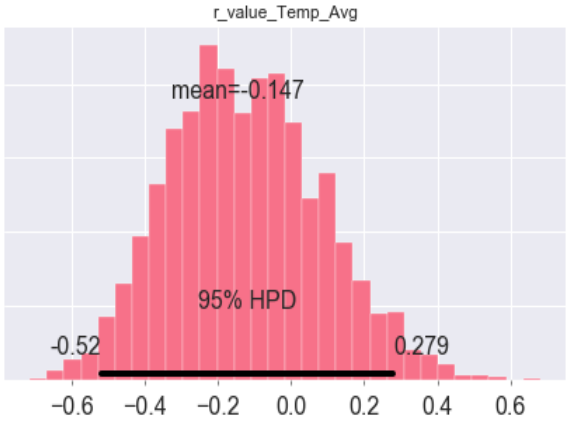


(b) 95 % de HPD.

Figura A.18: Datos de Taxi viejo.

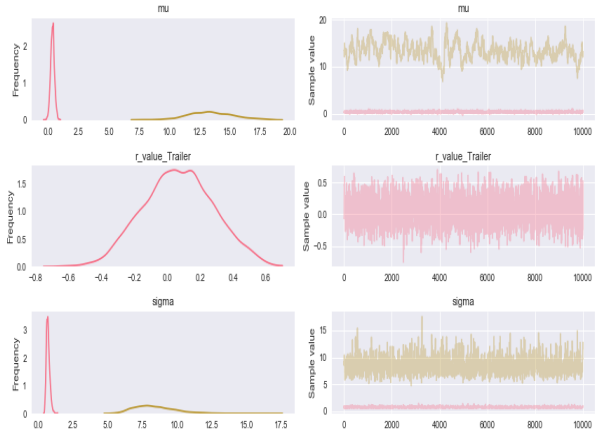


(a) *A posterioris* de los parámetros.

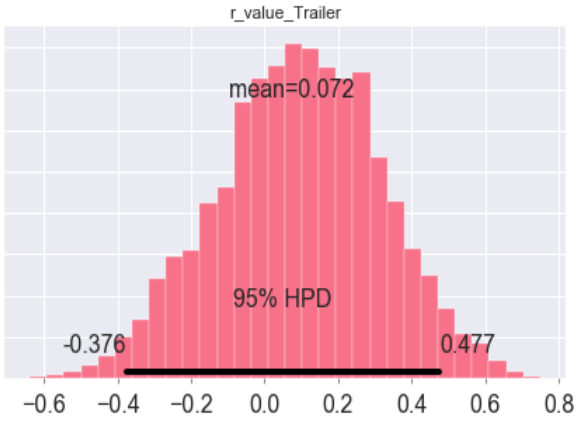


(b) 95 % de HPD.

Figura A.19: Datos de Temperatura.



(a) *A posterioris* de los parámetros.



(b) 95 % de HPD.

Figura A.20: Datos de Trailer.

Bibliografía

Ahrens, C. D. (2012). Meteorology today: an introduction to weather, climate, and the environment, ed. cengage learning.

Andrews, D. G. (2010). *An introduction to atmospheric physics*. Cambridge University Press.

Autores, V. (2017). Calidad del aire en la ciudad de méxico. informe 2016. urlhttp://www.aire.cdmx.gob.mx. Accedido 14-01-2019.

Beloconi, A., Chrysoulakis, N., Lyapustin, A., Utzinger, J., and Vounatsou, P. (2018). Bayesian geostatistical modelling of pm10 and pm2. 5 surface level concentrations in europe using high-resolution satellite-derived products. *Environment international*, 121:57–70.

Bernardo, J. M. and Smith, A. F. (2001). Bayesian theory.

Bolstad, W. M. and Curran, J. M. (2016). *Introduction to Bayesian statistics*. John Wiley & Sons.

Buhmann, M. (2017). Michael jd powell's work in approximation theory and optimisation.

Chen, L., Gao, S., Zhang, H., Sun, Y., Ma, Z., Vedal, S., Mao, J., and Bai, Z. (2018). Spatiotemporal modeling of pm 2.5 concentrations at the national scale combining land use regression and bayesian maximum entropy in china. *Environment international*, 116:300–307.

Chen, T.-M., Kuschner, W. G., Gokhale, J., and Shofer, S. (2007). Outdoor air pollution: nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. *The American journal of the medical sciences*, 333(4):249–256.

de Ecología y Cambio Climático, I. N. (2009a). El estado de la calidad del aire en méxico: 18 ciudades. urlhttp://www2.inecc.gob.mx/publicaciones2/libros/652/18ciudades.pdf. Accedido 25-01-2019.

de Ecología y Cambio Climático, I. N. (2009b). Guía metodológica para la estimación de emisiones vehículares. urlhttp://www2.inecc.gob.mx/publicaciones2/libros/618/vehiculos.pdf. Accedido 25-01-2019.

de Ecología y Cambio Climático, I. N. (2009c). Inventario de emisiones. urlhttp://www2.inecc.gob.mx/publicaciones2/libros/458/vehiculos.pdf. Accedido 25-01-2019.

Ghoumari, A., Nakib, A., and Siarry, P. (2018). Evolutionary algorithm with ensemble strategies based on maximum a posteriori for continuous optimization. *Information Sciences*, 460:1–22.

Gilbert, J. C. and Lemaréchal, C. (1989). Some numerical experiments with variable-storage quasi-newton algorithms. *Mathematical programming*, 45(1-3):407–435.

Gutiérrez Peña, E. (2013). El desarrollo de la estadística bayesiana. *Tema del mes*.

Hobbs, P. V. (2000). *Introduction to atmospheric chemistry*. Cambridge University Press.

Holton, J. R. and Hakim, G. J. (2012). *An introduction to dynamic meteorology*, volume 88. Academic press.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., and Brasseur, O. (2005). A neural network forecast for daily average pm 10 concentrations in belgium. *Atmospheric Environment*, 39(18):3279–3289.

Huang, S., Sun, Y., and Wu, Q. (2018). Stochastic economic dispatch with wind using versatile probability distribution and l-bfgs-b based dual decomposition. *IEEE Transactions on Power Systems*.

Hughes, M., Hall, A., and Fovell, R. G. (2007). Dynamical controls on the diurnal cycle of temperature in complex topography. *Climate dynamics*, 29(2-3):277–292.

Jonas, P. R. (1994). Back to basics: Why does it rain? *Weather*, 49(7):258–260.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., et al. (2003). Extensive evaluation of neural network models for the prediction of no 2 and pm 10 concentrations, compared with a deterministic modelling system and measurements in central helsinki. *Atmospheric Environment*, 37(32):4539–4550.

Le Tertre, A., Schwartz, J., and Touloumi, G. (2005). Empirical bayes and adjusted estimates approach to estimating the relation of mortality to exposure of pm10. *Risk Analysis: An International Journal*, 25(3):711–718.

Liou, K.-N. (2002). *An introduction to atmospheric radiation*, volume 84. Elsevier.

Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Malakoff, D. (1999). Bayes offers a 'new' way to make sense of numbers. *Science*, 286(5444):1460–1464.

OMS (2016). `http://www.who.int/mediacentre/factsheets/fs313/es/`, 22 de Noviembre de 2016.

Oprea, M., Dunea, D., and Liu, H.-Y. (2017). Development of a knowledge based system for analyzing particulate matter air pollution effects on human health. *Environmental Engineering & Management Journal (EEMJ)*, 16(3).

Pearce, D. and Crowards, T. (1996). Particulate matter and human health in the united kingdom. *Energy Policy*, 24(7):609–619.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55.

Seinfeld, J. H. and Pandis, S. N. (2016). *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons.

Sfetsos, A., Vlachogiannis, D., et al. (2010). Time series forecasting of hourly pm10 using localized linear models. *Journal of Software Engineering and Applications*, 3(04):374.

Silva, L. and Benavides, A. (2001). El enfoque bayesiano: otra manera de inferir. *Gaceta Sanitaria*, 15(4):341–346.

Strangeways, I. (2001). Back to basics: The 'met. enclosure': Part 6—wind. *Weather*, 56(5):154–161.

Stull, R. B. (2012). *An introduction to boundary layer meteorology*, volume 13. Springer Science & Business Media.

Tomasi, C., Fuzzi, S., and Kokhanovsky, A. (2017). *Atmospheric aerosols: life cycles and effects on air quality and climate*, volume 1. John Wiley & Sons.

Weber, S. A., Insaf, T. Z., Hall, E. S., Talbot, T. O., and Huff, A. K. (2016). Assessing the impact of fine particulate matter (pm2. 5) on respiratory-cardiovascular chronic diseases in the new york city metropolitan area using hierarchical bayesian model estimates. *Environmental research*, 151:399–409.

Wei-hua, A., Shu-rui, G., Hao, W., and Ming-bao, H. (2017). Planetary boundary layer height measured by a wind profiler based on the wavelet transform. *Journal of Tropical Meteorology*.

Wright, D. (1986). A note on the construction of highest posterior density intervals. *Applied statistics*, pages 49–53.

Yu, R., Liu, X. C., Larson, T., and Wang, Y. (2015). Coherent approach for modeling and now-casting hourly near-road black carbon concentrations in seattle, washington. *Transportation Research Part D: Transport and Environment*, 34:104–115.