ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

CONSTRUCCIÓN Y ESTIMACIÓN DE UN SISTEMA DE ÍNDICES DE PREVENCIÓN DE INCENDIOS FORESTALES (SIPIF) PARA EL PARQUE METROPOLITANO GUANGUILTAGUA DE QUITO

PROYECTO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

DIEGO FABIÁN MORALES NAVARRETE

diegofabian.morales@gmail.com

EDWIN ORLANDO QUIZHPI SÁNCHEZ

edorluan@hotmail.com

DIRECTOR: Dr. HOLGER ANÍBAL CAPA SANTOS, PhD.

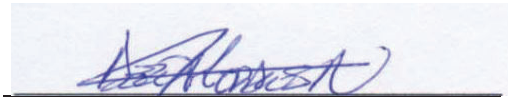
holger.capa@epn.edu.ec

QUITO, FEBRERO 2016

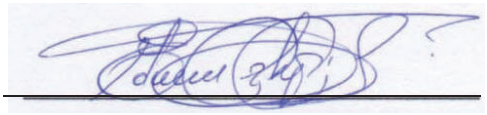
DECLARACIÓN

Nosotros, Diego Fabián Morales Navarrete y Edwin Orlando Quizhpi Sánchez, declaramos que el trabajo aquí descrito es de nuestra autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que hemos consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional, puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.



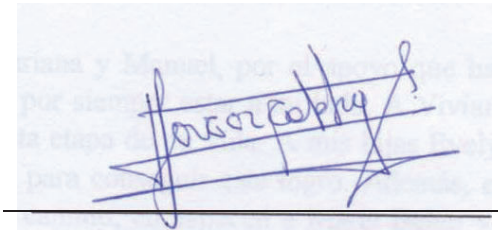
Diego Fabián Morales Navarrete



Edwin Orlando Quizhpi Sánchez

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Diego Fabián Morales Navarrete y Edwin Orlando Quizhpi Sánchez, bajo mi supervisión.



Dr. Holger Capa Santos, Ph.D.

DIRECTOR DE PROYECTO

AGRADECIMIENTOS

Tras un logro conseguido hay mucho más que el trabajo arduo de una persona, están sus creencias, sus familiares, su compañera de luchas, sus amigos; cada uno de ellos aportando de una manera especial y única.

Agradezco a Dios y a la Madre Dolorosa quienes han guiado mis pasos en todos los momentos de mi vida. A mis padres María y Hernán que han sido mi soporte y consejeros; quienes han sido la luz cuando el camino se tornaba oscuro. A mi hermano Andrés fuente inagotable de inspiración. A mi primo Daniel quien nunca dejo de confiar en mí. A mi familia que con su cariño me inculcaron que siempre hay que levantarse y seguir para conseguir algo más grande, no importan las veces que uno caiga. A Jacky quien no ha soltado mi mano ni en el momento más amargo y ha sido mi compañera de mis luchas. A Edwin que sin su incansable trabajo esto no se hubiera realizado. A mis amigos con quienes he disfrutado alegrías y tristezas.

Diego

Agradezco a mis padres, Mariana y Manuel, por el apoyo que han dado a mi vida. A mis hermanos, Mónica y Edison, por siempre estar a mi lado. A Viviana, que es la mujer que ha apoyado en gran medida a esta etapa de mi vida. A mis hijas Evelyn, Allison y Melanie, por ser la inspiración y la fuerza para conseguir este logro. Además, es importante reconocer el apoyo de los amigos en este camino, en especial a María Belén y a Jhosselyn por la ayuda brindada en las aulas de clases y fuera de ellas. A Diego ya que si él no se hubiera podido culminar con este proyecto.

Edwin

Al Ing. Enrique Palacios (EPMAPS-Q) por su desinteresada ayuda en el desarrollo de este proyecto. Finalmente, un agradecimiento especial al Dr. Holger Capa por el apoyo y guía durante este proyecto, por ser un maestro y amigo.

Diego, Edwin

DEDICATORIA

A Dios y a la Madre Dolorosa quienes me han acogido bajo su manto

A mis padres, Hernán y María, quienes han sido mi compañía en las noches de desvelo y me han enseñado que la humildad es la virtud más valiosa que se puede tener

A mi hermano Andres y mi primo Daniel, quienes con su apoyo nunca me dejaron desmayar

A mi familia, quienes me han inculcado valiosos valores

A mi novia Jacky, quien ha caminado hombro a hombro junto a mí en este camino y ha llenado de fortaleza mi vida

A mi padrino Pedro Fabián, quien me inculcó el luchar por un mundo mejor y que, aunque ya no este conmigo en la tierra, sigue vivo en mi corazón

Diego

A las mujeres de mi vida, Viviana, Allison, Evelyn, Melanie y Mariana.

Edwin

CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Los incendios forestales constituyen un problema de gran impacto social, económico y ambiental en el Distrito Metropolitano de Quito (DMQ); por ello, el presente proyecto desarrolla una metodología basada en la construcción de índices mediante herramientas estadísticas con un enfoque bayesiano para poder tener una aproximación sobre la ocurrencia de un incendio y la propagación del mismo; y más aún, poder proporcionar una aplicación que ayude en la predicción de estos fenómenos para que sea de utilidad en la planificación y, así, poder disminuir el impacto de los mismos.

Puesto que el DMQ posee una gran variedad de microclimas, el proyecto se enfoca en realizar un análisis sobre un sitio de gran afectación: el parque metropolitano Guangüiltagua. Uno de los inconvenientes que se tuvieron dentro del desarrollo del proyecto fue la mala calidad de las bases de datos sobre incendios y la poca cantidad de información recabada dentro del parque metropolitano y, aunque las herramientas bayesianas ayudan a mitigar la falta de información, se realizan modelos en los cuales la información de incendios es procedente de todo el DMQ pero bajo las condiciones meteorológicas del parque metropolitano, con ello se logra obtener los índices propuestos.

El índice de prevención de incendios, provee una probabilidad asociada a la ocurrencia de un incendio y ell índice de propagación de incendios, proyecta el número de posibles hectáreas consumidas por el fuego si un incendio llegara a producirse. Los índices se determinan por 4 categorías que van desde *bajo* hasta *grave.*

Con estos dos índices se constituye el Sistema de Prevención y Propagación de Incendios Forestales (SIPIF), que es el objetivo de este proyecto. Adicionalmente, los resultados se muestran a través de un aplicativo web desarrollado en el módulo *Shiny* del paquete estadístico *R Project*.

Palabras claves: Incendios, estadística bayesiana, modelos logístico bayesiano, prevención y propagación de incendios, componentes principales bayesianas, modelos lineales bayesianos, índice de prevención de incendios, índice de propagación de incendios, R Project, Shiny.

ABSTRACT

A huge social, economical and environmental impact problem at the Distrito Metropolitano de Quito (DMQ) are the forest fires. Therefore, in this project we developed a methodology to estimate wildfires occurrence and their spread. It is based on the construction of indexes using Bayesian statistical tools. Moreover, we provided an application for forest fires prediction which can be applied as a tool for fire-fighting planning and thus reducing their impact.

Since the DMQ has a variety of micro-climates, this project was focused on the analysis of a highly wildfire-affected area: Guangüiltagua Metropolitan Park. In this context, the most critical drawbacks we found were the poor quality of the forest fires databases and the lack of information regarding the Metropolitan Park. Even though Bayesian tools help to mitigate this lack of information, it was necessary to develop models where the information was extracted from the whole DMQ forest fires database but considering only the data under the same weather conditions of the Metropolitan Park. Altogether, this approach allowed us to obtain the proposed rates.

Whereas the fire prevention index provides the probability associated with the fire occurrence, the fire spread index maps out the probable number hectares consumed by the fire if a wildfire occurs. Both indexes were defined by using four categories ranging from low to severe.

The main goal of this project: the Prevention System and Forest Fire Propagation (PSFF) was constituted by the above mentioned indexes. Additionally, the results are displayed through a web application developed using the Shiny module, part of the statistical package R-Project.

Keywords: Forest fires, Bayesian statistics, Bayesian logistic models, prevention and spread of wildfires, Bayesian principal components, Bayesian linear models, fire prevention index, fire spread index, R-Project, Shiny.

CAPÍTULO 1
INTRODUCCIÓN

Desde hace muchos años en Ecuador, los incendios forestales son parte de la convivencia de los habitantes en los meses de verano o periodo seco en el país. En la ciudad de Quito, este fenómeno es muy recurrente sobre todo en los meses de junio a septiembre de cada año y representa un importante impacto en los recursos humanos, económicos, ecológicos y ambientales para combatir este fenómeno. Así, por ejemplo, en el período comprendido entre el 26 de junio y el 7 de septiembre del 2014, se registraron un total de 1.081 incidentes por fuego, con 526,04 Ha. afectadas[1].

Puesto que el DMQ es una extensión territorial grande y la posibilidad de obtener información sobre variables ambientales, meteorológicas y topográficas es limitada, en este proyecto se realiza el análisis y la construcción de un sistema de índices para la prevención y propagación de los incendios forestales en el Parque Metropolitano Guangüiltagua, el que está localizado en la loma de Guangüiltagua, al norte de la ciudad de Quito y tiene una extensión de 557 Ha., constituyéndose en el principal pulmón de la ciudad y uno de los lugares afectado por los incendios. En el 2014 el número de incendios que ocurrieron fue de 7 con un total de 2,09 Ha. de afectación.

Dadas estas afectaciones, es importante tener un plan de prevención de incendios para evitar todos los problemas causados por ellos. En este sentido, se propone implementar un sistema de índices cuantitativos que permitan determinar la probabilidad de que un incendio suceda, la cantidad de hectáreas que consumiría y el tiempo que le tomaría en hacerlo.

En el DMQ, el Comité de Operaciones y Emergencias (COE-Q) actualmente utiliza un índice de propagación de incendios forestales basado en la metodología desarrollada por Rodríguez y Moretti (1988), el que se origina en una análisis de correlaciones entre variables meteorológicas (temperatura, humedad relativa, velocidad del viento y días consecutivos de sequía) y la ocurrencia y magnitud de los incendios en los periodos comprendidos entre los

[1] COE, "Informe de situación: Plan de prevención y respuesta para incendios forestales del DMQ, Período: 26 de junio a 7 de septiembre de 2014".

años 1984 y 1987, que sucedieron en la región Andino- Patagónica. El valor del índice se obtiene sumando los valores de las tablas Temperatura, Humedad, Velocidad del Viento y Días consecutivos de sequía de la Tabla 1, en las que las variables ingresadas se obtienen de observaciones efectuadas a las 15 horas. Los días secos o sin presencia de lluvia se cuentan a partir del último día con precipitación menor de 2 mm. Cuando se hace el cálculo en un día que se produce precipitación, el valor a sumar en la tabla *Días consecutivos de sequía* es de 0; al segundo día con precipitación, el valor obtenido de la suma de los valores de las tablas Temperatura, Humedad y Velocidad del viento se multiplica por un factor de corrección, en este caso 0,8; al tercer día de lluvia se multiplica por 0,6; al cuarto día de lluvia por 0,4, al quinto día de lluvia por 0,2 y a partir del sexto día de lluvia los coeficientes se multiplican por 0 (el valor del índice es 0). Por ejemplo, si se considera:

- Temperatura: 25°, Humedad 60%, Velocidad del viento: 12 km/h y Días secos:0 (con un día de lluvia), se tendría: $I = 22{,}5 + 12{,}5 + 7{,}5 + 0 = 42{,}5$

- Temperatura: 25°, Humedad 60%, Velocidad del viento: 12 km/h y Días secos:0 (con dos días de lluvia), se tendría: $I = 0{,}8 * (22{,}5 + 12{,}5 + 7{,}5 + 0) = 34$

- Temperatura: 25°, Humedad 60%, Velocidad del viento: 12 km/h y Días secos:0 (con cinco días de lluvia), se tendría: $I = 0{,}2 * (22{,}5 + 12{,}5 + 7{,}5 + 0) = 8{,}5$

- Temperatura: 25°, Humedad 60%, Velocidad del viento: 12 km/h y Días secos:0 (con seis días de lluvia o más), se tendría: $I = 0 * (22{,}5 + 12{,}5 + 7{,}5 + 0) = 0$

Tabla 1. Puntaje de variables dentro del índice de Rodríguez y Moretti

Temperatura (°C)	Índice
menos de 10	2,5
10 a 11,9	5,0
12 a 13,9	7,5
14 a 15,9	10,5
16 a 17,9	12,0
18 a 19,9	15,5
20 a 21,9	17,5
22 a 23,9	20,0
24 a 25,9	22,5
26 o más	25,0

Humedad (%)	Índice
80 o más	2,5
79,9 a 75	5,0
74,9 a 70	7,5
69,9 a 65	10,5
64,9 a 60	12,5
59,9 a 55	15,0
54,9 a 50	17,5
49,9 a 45	20,0
44,9 a 40	22,5
39,9 o menos	25,0

Velocidad del viento (km/h)	Índice
menos de 3,0	1,5
3,0 a 5,9	3,0
6,0 a 8,9	4,5
9,0 a 11,9	6,0
12,0 a 14,9	7,5
15,0 a 17,9	9,0
18,0 a 20,9	10,5
21,0 a 23,9	12,0
24,0 a 26,9	13,5
27,0 o más	15,0

Dias consecutivos de sequía	Índice
1	3,5
2 a 4	7,0
5 a 7	10,5
8 a 10	14,0
11 a 13	17,5
14 a 16	21,0
17 a 19	24,5
20 a 22	28,0
23 a 25	31,5
26 en más	35,0

Fuente: Evaluación de peligro de incendios. Informes técnicos. Informe Técnico Nº1. Sistemas de Evaluación de Peligro de Incendios.
Elaboración: Autores

El rango del índice varía de 0 a 100 y está dividido en cuatro clases que indican el grado de peligro de propagación si ocurriera un fuego, tal como se observa a continuación:

Tabla 2. Calificación de riesgo para el índice de Rodríguez y Moretti

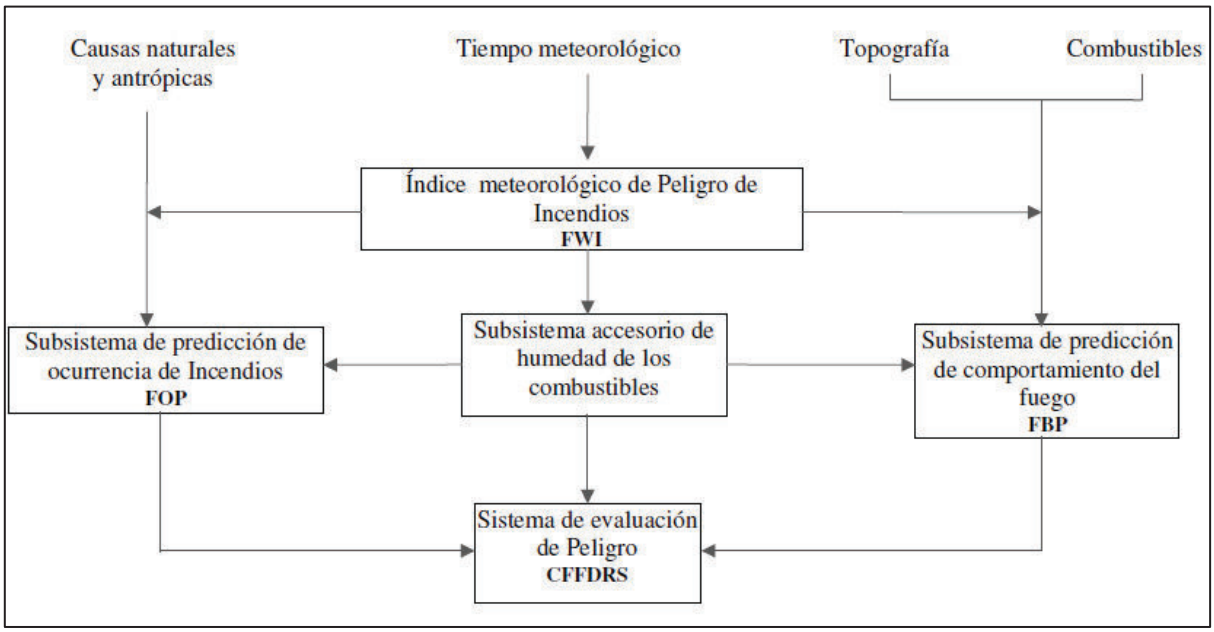
Rango	Calificación
0,0 – 24,0	Leve
24,1 – 49,0	Moderado
49,1 – 74,0	Alto
75,1 – 100,0	Extremo

Fuente: Evaluación de peligro de incendios. Informes técnicos. Informe Técnico Nº1. Sistemas de Evaluación de Peligro de Incendios.
Elaboración: Autores

Sin embargo, este índice no es lo suficientemente robusto, ya que por estudios realizados en varios lugares del mundo, la ignición y propagación de los incendios también depende de variables topográficas y de variables ambientales (combustibles vegetales) presentes al momento de producirse el fuego[2,3]. En este sentido, un modelo desarrollado en Canadá (Sistema de evaluación de peligro de incendios canadiense) calcula un índice derivado de los efectos combinados de las siguientes componentes: predicción de ocurrencia de incendios, índice de propagación y de combustible disponible, y códigos de humedad de combustible; la estructura del mismo está dada en la siguiente figura:

[2] Dirección General de Protección Civil y Emergencias –Ministerio del Interior-España. Red Radio Emergencia – VADEMECUM REMER
[3] Canadian Forest Fire Danger Rating System (CFFDRS)- http://cwfis.cfs.nrcan.gc.ca/background/summary/fdr

Figura 1. Estructura del sistema de índices de prevención de incendios desarrollado por CFFDRS



Fuente: Evaluación de peligro de incendios. Informes técnicos. Informe Técnico Nº1. Sistemas de Evaluación de Peligro de Incendios.

Para la investigación que se propone realizar se utilizan parámetros similares a los utilizados en Canadá como un punto de partida, y técnicas de estimación bayesiana para la construcción del sistema propuesto.

La estructura del presente estudio es la siguiente: En el capítulo 2, se abordan los aspectos teóricos necesarios para comprender la metodología utilizada en la construcción y estimación del sistema de índices para la prevención y propagación de los incendios forestales; este abarca desde la comprensión de los factores necesarios para que un incendio ocurra, así como las técnicas estadísticas necesarias basadas en la estimación bayesiana. En el capítulo 3 se describe la metodología y desarrollo de la estimación del sistema de índices con la información disponible. En el capítulo 4 se presentan los resultados obtenidos y la implementación en un mapa que muestre gráficamente dichos resultados. Finalmente, en el capítulo 5 se presentan las conclusiones y recomendaciones encontradas durante el desarrollo de este proyecto, luego de analizar el sistema de índices en diferentes escenarios.

CAPÍTULO 2
MARCO TEÓRICO

En este capítulo se describen las nociones y conceptos necesarios para comprender la metodología utilizada en la construcción y estimación del sistema de índices para prevenir los incendios en el parque Metropolitano Guangüiltagua.

2.1 INCENDIOS FORESTALES

Un incendio forestal es un fenómeno que se produce cuando se aplica suficiente calor a un combustible vegetal situado en un terreno forestal.

2.1.1 CARACTERÍSTICAS DE LOS INCENDIOS FORESTALES

El fuego es un proceso químico que ocurre cuando se aplica una fuente de calor a una sustancia **combustible.** Para que se realice la combustión es necesario se combinen: oxigeno, gases que desprenden del combustible y suficiente nivel de energía. Esto producirá calor y luz.

En el caso de los incendios forestales, el fuego es una reacción exotérmica que se mantendrá mientras haya combustible y las condiciones en las que se inició no cambien. El combustible en un incendio forestal es la masa vegetal del bosque y en un bosque el oxígeno no faltará.

En un bosque se encuentran diferentes tipos de combustibles. Según el tipo de combustible la reacción puede ser más fuerte y espontánea a temperatura ambiente:

- **Combustibles vivos:** Plantas vivas con un contenido hídrico elevado que depende de las lluvias y sequías, especie vegetal, tipo de suelo, exposición, etc.

- **Combustibles muertos:**

Ligeros: Fácilmente inflamables y de rápida propagación que se secan fácilmente.

Pesados: Se secan lentamente y son propensos a ser quemados después de una larga sequía.

En el caso de esta investigación, el parque Guangüiltagua tiene todos los factores para que se produzca un incendio forestal: gran cantidad de oxígeno (se encuentra en una loma), en época seca combustibles vivos con muy poco contenido hídrico y combustibles muertos (ligeros y pesados).

2.1.2 TIPOS DE INCENDIOS FORESTALES

Se diferencian tres tipos de incendios forestales según donde se localicen dentro del bosque:

- **Incendio superficial:** Este es el tipo de incendio más frecuente en el mundo; quema hierbas, matorrales secos y restos vegetales sobre el suelo.

- **Incendio de copas:** En este tipo de incendios se queman las copas de los árboles; se necesita que exista una pendiente muy pronunciada, la densidad arbórea sea alta y una gran cantidad de viento.

- **Incendio subterráneo:** Este tipo de incendios son muy poco frecuentes en el mundo; queman la capa de materia orgánica acumulada en el suelo y las raíces de los árboles que encuentren.

2.1.3 CAUSAS Y FACTORES AGRAVANTES

La propagación de un incendio forestal obedece a un mecanismo complejo; es decir, se puede quemar un árbol de manera aislada sin que el fuego se propague o puede suceder que el fuego encienda los árboles alrededor y, entonces, el incendio crezca. Los factores que se pueden considerar en el inicio y propagación de un incendio son:

- **Combustible:** Como ya se ha expuesto anteriormente, dependiendo del tipo de vegetación que se encuentre en el área del incendio y su grado de humedad, es más fácil o no que un incendio se inicie. Además, la cantidad y continuidad de vegetación (horizontal y vertical) favorecen a la propagación de un incendio.

- **Relieve:** El fuego se propaga de manera diferente en una parte plana que en una pendiente; mientras más pendientes en el terreno el incendio se propaga de manera más rápida. Además, el relieve accidentado dificulta las tareas de control de un incendio.

- **Meteorología:** La propagación del incendio es más fácil cuando el viento es intenso y constante, la temperatura es alta y la humedad ambiental es baja. La radiación solar es importante, tanto en el inicio como en la propagación de un incendio.

El inicio del fuego puede deberse a tres posibles causas:

i. **Natural:** Cuando un incendio se inicia por causas naturales puede deberse a temperaturas extremadamente altas (muy raro) y que exista vegetación de fácil ignición; también por, una tormenta eléctrica, erupción volcánica, etc.

ii. **Antrópico:** En muchos lugares del mundo y en particular en el Ecuador, la mayoría de incendios son provocados de manera intencionada o inintencionada por personas:

    - Intencionado: Provocado por pirómanos (personas con problemas sociales o psíquicos, en su mayoría).
    - Negligencia: Provocado por personas sin intención de hacerlo; por ejemplo, botar una colilla de cigarrillo al lado de una carretera con vegetación muerta y seca.
    - Accidentes: Provocados por un accidente de tránsito terrestre o aéreo.
    - Otros: Pueden ser provocados por fuegos controlados que se escapan, como trozos de vidrio que cumplen una función de lupa.

iii. **Desconocido:** Pese a los esfuerzos por conocer las causas de un incendio, en un importante porcentaje de ellos no se puede determinar la causa.

2.1.4 MEDICIÓN Y EVALUACIÓN

Un incendio forestal se desarrolla a partir de un ***frente de avance***, que es una línea irregular que se desplaza y va quemando a medida que encuentra combustible. Delante de esta línea existe el ***frente de desecamiento***; este es invisible, sus altas temperaturas secan los vegetales y los matan preparándolos para ser quemados. Las dimensiones de estos frentes permiten evaluar la ***magnitud*** del incendio.

La ***intensidad*** de un incendio se refiere a la temperatura que alcanza y la velocidad con la que se propaga. En general, la capacidad destructiva del incendio aumenta con la temperatura y disminuye con la velocidad de propagación.

2.2 CONCEPTOS DE ESTADÍSTICA BAYESIANA

El objetivo de los métodos y procedimientos desarrollados por la estadística y la estadística Bayesiana, es proporcionar una metodología para analizar de manera adecuada un conjunto de datos de manera que se convierta en información útil.

En muchos aspectos de la vida se utiliza el concepto de probabilidad (aunque a veces no se conoce exactamente a qué se refiere). Dentro del estudio de la estadística y la probabilidad existen por lo menos tres visiones diferentes de la interpretación de la probabilidad:

- ***Clásica:*** Supone que un experimento aleatorio produce resultados igualmente verosímiles (posibles) y se calcula como el cociente entre los casos favorables y los posibles.

Sea A un evento que ocurre n veces (n_A); entonces:

$$P(A) = \frac{n_A}{n} \tag{2.1}$$

- ***Frecuentista:*** Supone que un experimento aleatorio puede ser repetido un número infinito de veces bajo condiciones similares y propone como medida de probabilidad a la proporción de las veces que ocurre el evento de interés.

  Sea A un evento que ocurre un número infinito de veces:

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} \tag{2.2}$$

- ***Subjetiva:*** Es simplemente una medida de la incertidumbre, asociada a un evento, asignada por un decisor. En otras palabras, es un juicio particular (del analista) sobre la verosimilitud de que ocurra un evento.

El análisis bayesiano se basa en la interpretación subjetiva de la probabilidad y tiene como punto central el teorema de Bayes (se expondrá más adelante); este proporciona un modo natural de actualización de las creencias cuando aparece nueva información. Es decir, el análisis bayesiano es un proceso de aprendizaje deductivo.

La metodología del análisis bayesiano aplicable a cualquier técnica estadística, tiene tres etapas fundamentales:

1. Especificar un modelo de probabilidad que incluya algún tipo de conocimiento (*a priori*) sobre los parámetros del modelo a estimar.
2. Actualizar el conocimiento sobres los parámetros desconocidos condicionando este modelo a los datos observados.
3. Evaluar el ajuste del modelo (*a posteriori*) a los datos y la sensibilidad de las conclusiones a cambios de los supuestos del modelo.

Se puede decir que la inferencia bayesiana tiene una gran similitud con la interpretación clásica de la probabilidad; es decir, existe un parámetro poblacional que se desea estimar a partir de una muestra. Sin embargo, la diferencia fundamental radica en que la inferencia bayesiana considera el parámetro a estimar como una ***variable aleatoria***.

2.2.1 DEFINICIONES Y TEOREMAS BÁSICOS

En la estadística bayesiana es muy importante el concepto de la ***probabilidad condicional***:

***Definición*:** Sean dos eventos, A y B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.3}$$

Esta definición se aplica tanto a variables discretas como a variables continuas.

Desde el punto de vista bayesiano, todas las probabilidades son condicionales porque casi siempre existe algún conocimiento previo acerca de los eventos.

Otro concepto que es de gran ayuda en el análisis bayesiano es el de ***probabilidad total:***

***Definición*:** Sea un evento A y una partición de eventos $B_1, \dots B_n$:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)\, P(B_i) \tag{2.4}$$

Se puede aplicar a variables discretas y variables continuas.

2.2.1.1 Teorema de Bayes

Sean $A_1, A_2, \dots, A_n$ eventos incompatibles dos a dos, tal que $\Omega = \bigcup_i A_i$ (Ω es el espacio muestral). Sea un evento B, tal que $P(B) > 0$. Se suponen conocidas tanto las $P(B|A_i)$ como las $P(A_i)$. El problema de Bayes consiste en calcular, con los datos anteriores, las probabilidades $P(A_k|B), k = 1,2, \dots n$. Se tiene:

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)\, P(A_i)} \tag{2.5}$$

2.2.1.2 Distribución de probabilidad a posteriori

Dentro de las siguientes secciones de este capítulo, y en algunas otras de los capítulos siguientes, se utiliza por facilidad y abuso del lenguaje a X como una variable aleatoria y a x como una observación de la variable aleatoria X o, a veces, como una muestra de X.

Sea $x = (x_1, x_2, \dots, x_n)'$ un vector aleatorio de datos observados y sea θ un vector de parámetros desconocidos. Se considera la función de probabilidad o densidad $f(x|\theta)$; se supone que θ tiene una distribución de probabilidad *a priori* $\pi(\theta)$. La inferencia concerniente a θ está basada en su distribución ***a posteriori***, dada por el cociente de la distribución conjunta de x con respecto a θ $[h(x,\theta)]$ y la distribución marginal de x $[m(x)]$.

$$\pi(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{h(x,\theta)}{\int h(x,\theta)d\theta} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \tag{2.6}$$

Como se puede observar la inferencia se basa en la distribución de probabilidad del parámetro (desconocido) dados los datos observados, en lugar de encontrar la distribución de los datos dado el valor del parámetro.

2.2.1.3 Función de verosimilitud

La función de verosimilitud presenta el proceso a través del cual aparecen las variables x en términos del parámetro desconocido θ.

***Definición*:** Sea una muestra aleatoria de tamaño n: $X_1, X_2, \dots, X_n$, de la variable aleatoria X; se define a la función de verosimilitud como:

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2.7}$$

Desde el punto de vista bayesiano, la función de verosimilitud mide el grado de "creencia" del investigador de que los datos tomen ciertos valores dada la información hipotética de que los parámetros tengan cierta distribución, además de toda la información *a priori*.

El principio de verosimilitud

Un principio muy importante para el paradigma bayesiano es el de verosimilitud, ya que permite establecer que con una muestra de la variable aleatoria X, debería ser importante para tener evidencia (o incluso conclusiones) sobre θ.

Existe una gran cantidad de aproximaciones para obtener inferencias mediante la función de verosimilitud. Una de estas aproximaciones (en extremo) es el principio de verosimilitud:

PRINCIPIO DE VEROSIMILITUD: Si para dos modelos, la combinación de datos conduce a funciones de verosimilitud equivalentes, las inferencias sobre el parámetro desconocido deben ser idénticas.

Es importante considerar algunos puntos con respecto a este principio:

- La correspondencia de información a partir de dos funciones de verosimilitud, se aplica solo en el caso que se refieran al mismo parámetro.
- Para los métodos bayesianos la verosimilitud es un instrumento para pasar de la distribución *a priori* (se definirá más adelante) a la distribución *a posteriori.*
- Es muy importante la interpretación correcta de $L(\theta|x)$. Esta es una función de θ, dado x, a través de la cual se pueden realizar diferentes tipos de inferencias, sea en el campo bayesiano como en el no bayesiano.
- El interés de la función de verosimilitud no es el valor en sí, sino la *razón de verosimilitud* la que nos informa la confianza para cada valor de θ.
- El principio de verosimiltud se cumple si:

$$L_1(\theta|x) = cL_2(\theta|x) \quad \forall x, \forall \theta, c \text{ una constante}$$

Es decir, la razón de verosimilitud para dos experimentos y cada observación, es constante.

2.2.2 PROBABILIDAD SUBJETIVA *"A PRIORI"*

Las creencias se pueden expresar en términos de apuestas; estas deben ser elaboradas de tal manera que una persona no tenga certeza de ganancia o pérdida. Esta condición obliga al individuo a asignar las apuestas de acuerdo a sus creencias.

Además, una vez que se ha fijado la posibilidad de ganar o perder, el individuo está preparado para apostar en cualquier dirección; además, siendo moral y honesto, permite que las reglas básicas de la probabilidad sean derivadas como teoremas.

Las distribuciones a priori se clasifican en:

- Distribuciones a priori propias e impropias
- Distribuciones a priori informativas y no informativas
- Distribuciones a priori conjugadas y no conjugadas

2.2.2.1 Distribución a priori propia

Esta distribución asigna pesos no negativos y que se suman o integran (si son variables discretas o continuas, respectivamente) hasta uno, a todos los valores posibles del parámetro. Así, una distribución propia satisface las condiciones de función de densidad de probabilidad.

2.2.2.2 Distribución a priori impropia

Es una distribución que suma o integra (si son variables discretas o continuas, respectivamente) a un valor diferente de uno, notado por $\boldsymbol{M}$. Si $\boldsymbol{M}$ es finito, entonces la distribución impropia induce una distribución propia normalizando la función. Si $\boldsymbol{M}$ es infinito, entonces la distribución tiene un papel de ponderación o de herramienta técnica para llegar a una distribución *a posteriori*.

2.2.2.3 Distribución a priori no informativa

Se dice que una distribución a priori es no informativa cuando refleja una ignorancia total o un conocimiento muy limitado sobre el parámetro de interés; o, se desea que los "datos hablen por ellos mismos". Este es un campo de estudio que ha crecido enormemente; además, este tipo de distribuciones tiene una importancia especial ya que, en general, se conoce muy poco o nada sobre el parámetro de interés.

Más adelante se realizará una explicación más detallada de estas distribuciones.

2.2.2.4 Distribución a priori informativa

Una distribución a priori informativa es aquella que refleja en su totalidad la información del parámetro de interés, y por ello mismo pierde sentido realizar inferencias estadísticas sobre la misma.

2.2.2.5 Distribución a priori conjugada

Se dice que una distribución a priori es conjugada cuando esta coincide con la distribución *a posteriori* del parámetro, excepto en los hiperparámetros (parámetros pertenecientes a las distribuciones a priori).

2.2.2.6 Consistencia Posterior

***Definición*:** La distribución *a posteriori* se dice que es consistente en un valor dado θ_0, si para cualquier vecindad V de θ_0, $\pi(\theta_n \notin V|x) \to 0$ (en probabilidad) cuando $n \to \infty$ donde θ_0 es el verdadero valor del parámetro.

Es decir, en general, no importa la distribución *a priori* de los datos si el tamaño muestral crece indefinidamente. Lo que es fundamental es el experimento que genera los datos, que se presupone *insesgado*.

2.2.2.7 Suficiencia

Suponga que la distribución de una variable aleatoria X depende de un parámetro desconocido θ. Un estadístico $T(X)$ es *suficiente* si la distribución condicional de X dada por $T(X) = t$, no depende de θ.

***Lemma (Criterio de factorización de Neyman)*:** Suponga que se tiene la variable aleatoria X con función de distribución conjunta $f(x|\theta)$. Luego, un estadístico $T = T(X)$ es suficiente para θ si y solamente si,

$$f(\theta|x) = f(x)g(T(x), \theta)$$

donde,

f: es una función que depende de X.

g: es una función que depende del estadístico $T(x)$ y del parámetro θ.

La función de verosimilitud (como un estadístico suficiente) se puede factorizar como:

$$L(\theta|x) = f(x|\theta) = f(x)g(T(x), \theta) \tag{2.8}$$

***Principio de suficiencia*.** Sean dos observaciones diferentes $x\ e\ y$, tal que tienen el mismo valor de estadístico $T(x) = T(y)$, de un estadístico suficiente de familia $f(.|\theta)$. Luego, las inferencias sobre θ basados en $x\ e\ y$ deben ser las mismas (Knight, 2000).

2.2.3 DISTRIBUCIÓN A PRIORI NO INFORMATIVA

Una distribución sobre θ se dice que es no informativa si no contiene información sobre θ. Por ejemplo, considerando una distribución binomial se puede definir una distribución no informativa asignándole el valor de $p = \theta = 0{,}5$.

2.2.3.1 Método de Jeffreys

Caso univariante

Antes de describir el método de Jeffreys es necesario recordar la siguiente definición:

***Definición*:** Si $\theta \in \mathbb{R}$, se define la información esperada de Fisher como:

$$I(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log\big(f(x|\theta)\big) \right] \tag{2.9}$$

donde,

E_θ: El valor esperado condicional de la variable aleatoria que es función de x.
$f(x|\theta)$: Función de densidad de x (que depende de θ).

Si un investigador no tiene conocimiento con respecto a un parámetro θ, entonces su opinión acerca de θ dada las observaciones de la variable aleatoria X, debe ser la misma que el de una parametrización para θ o cualquier función inyectiva de θ (se notará por $h(\theta)$ a una función inyectiva de θ).

Si la distribución a priori no informativa sobre θ es $\pi(\theta)$, la a priori no informativa sobre $h(\theta)$ debe ser

$$\pi\big(h^{-1}\big(h(\theta)\big)\big) \left| \frac{dh^{-1}}{dh(\theta)}\big(h(\theta)\big) \right| \tag{2.10}$$

Jeffreys propuso una solución a este problema, definiendo la distribución a priori de manera proporcional a la raíz cuadrada de la información esperada de Fisher; es decir:

$$\pi(\theta) \propto |I(\theta)|^{\frac{1}{2}}. \tag{2.11}$$

Notación: El símbolo "$\propto$" significa "es proporcional a"; por ejemplo, $A \propto B$ se lee como: "A es proporcional a B".

Caso multivariante

***Definición*:** Si $\theta \in R^n$, se define la matriz de información de Fisher como aquella matriz nxn cuyas componentes son:

$$I_{ij}(\theta) = -E_\theta \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log\big(f(x|\theta)\big) \right] \tag{2.12}$$

Jeffrey propone que se elija como función a priori no informativa la determinada por la siguiente expresión:

$$\pi(\theta) \propto \left[\det I_{ij}(\theta)\right]^{\frac{1}{2}} \tag{2.13}$$

2.2.4 DISTRIBUCIÓN A POSTERIORI

La información a posteriori de θ dado x, con $\theta \epsilon \Theta$, está dada por la expresión $\pi(\theta|x)$, que se conoce después de observar los datos x. Entonces, la función de densidad (subjetiva) conjunta se define por:

$$h(x, \theta) = \pi(\theta) L(\theta|x) \tag{2.14}$$

donde,

$\pi(\theta)$: densidad a priori de θ.

$l(\theta|x)$: función de verosimilitud.

Además, x tiene la función marginal dada por:

$$m(x) = \int L(\theta|x)\pi(\theta)d\theta \tag{2.15}$$

Si la función marginal $m(x) \neq 0$, entonces:

$$\pi(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{\pi(\theta)L(\theta|x)}{\int L(\theta|x)\pi(\theta)d\theta} \tag{2.16}$$

Por otro lado, si se considera un estadístico suficiente de θ, $T(x)$ [es decir, la función de verosimilitud se puede factorar como: $L(\theta|x) = f(x)g(T(x),\theta)$], considerando $m(x) \neq 0$ la densidad marginal para $T(x) = t$, se cumple que:

$$\pi(\theta|x) = \pi(\theta|t) = \frac{g(t|\theta)\pi(\theta)}{m(t)} \tag{2.17}$$

De esta relación se puede concluir que la distribución a posteriori es proporcional a la función de verosimilitud multiplicada por la distribución a priori.

2.2.5 INFERENCIA BAYESIANA

Dado que la distribución a posteriori contiene toda la información disponible acerca del parámetro θ, algunas inferencias pueden ser parte únicamente de las características de esta distribución. A continuación, se presentan algunas técnicas que son de utilidad

2.2.5.1 Estimación puntual

Se denota como $\hat{\theta}$ al estimador de θ. $\hat{\theta}$ es el valor que maximiza la función de verosimilitud $L(\theta|x)$.

***Definición*:** La estimación de máxima verosimilitud generalizada de θ es la moda más grande $\hat{\theta}$ de $\pi(\theta|x)$. Es decir, el valor $\hat{\theta}$ que maximiza $\pi(\theta|x)$, considerada función de θ.

Otro estimador bayesiano común de θ es la media de la distribución a posteriori $\pi(\theta|x)$. En el contexto de la teoría de la decisión, se puede demostrar que utilizando una función de pérdida cuadrática (que es la más utilizada por su simplicidad y buenas propiedades matemáticas), el estimador óptimo de θ es la esperanza de la distribución a posteriori.

2.2.5.2 Error de estimación

Como en la mayoría de procedimientos estadísticos, siempre que se realiza una estimación está inmerso un error, el cual debe ser lo más pequeño posible. Por lo tanto, es importante siempre tener una medida de dicho error. La medida bayesiana que se utiliza para medir la precisión de una estimación univariante (error) es la varianza a posteriori de la estimación.

***Definición*:** Sea θ un parámetro de valor real con distribución *a posteriori* $\pi(\theta|x)$ y sea λ *a priori* el estimador de θ, se define la varianza a posteriori de λ como:

$$V_{\lambda}^{\pi} = E^{\pi(\theta|x)}[(\theta - \lambda)^2] \tag{2.18}$$

Notación: Se nota la varianza *a posteriori* como: V_{λ}^{π}, donde el superíndice π hace referencia a la distribución *a posteriori* y el subíndice λ hace referencia al estimador de θ.

Cabe recalcar que al tener un estimador $\hat{\theta}$ de θ, en el caso de la definición λ, reemplazando λ se obtendría la varianza *a posteriori.*

Cuando λ es la media *a posteriori*; es decir:

$$\lambda = \mu^{\pi}(x) = E^{\pi(\theta|x)}[\theta] \tag{2.19}$$

Notación: Se nota la media *a posteriori* como: μ^{π}, donde el superíndice π hace referencia a la distribución *a posteriori*.

Entonces,

$$V^{\pi}(x) = V_{\mu}^{\pi}\pi(x) \tag{2.20}$$

$V^{\pi}(x)$, se denomina varianza a posteriori [en efecto es la varianza de θ para la distribución $\pi(\theta|x)$]. La desviación estándar a posteriori es la raíz cuadrada de la varianza a posteriori presentada anteriormente.

Se puede verificar que:

$$V_{\lambda}^{\pi} = V^{\pi}(x) + (\mu^{\pi} - \lambda)^2 \tag{2.21}$$

El caso multivariante

Sea un vector de parámetros $\theta = (\theta_1, \dots \theta_n)'$, la estimación de máxima verosimilitud generalizada (moda a posteriori) es $\hat{\theta} = (\hat{\theta}_1, \dots \hat{\theta}_n)'$. Sin embargo, dado que la existencia y unicidad del estimador no se garantiza en el caso multivariante, es mejor utilizar la media a posteriori; es decir:

$$\mu^{\pi}(x) = \left(\mu_1^{\pi}(x), \dots, \mu_n^{\pi}(x)\right)' = E^{\pi(\theta|x)}[\theta] \tag{2.22}$$

La varianza del error viene expresada por:

$$V^{\pi}(x) = E^{\pi(\theta|x)}\left[\left(\theta - \mu^{\pi}(x)\right)\left(\theta - \mu^{\pi}(x)\right)'\right] \tag{2.23}$$

En este caso se la varianza corresponde a la matriz de varianzas-covarianzas. Por lo tanto, el error estándar de la estimación $\mu_i^{\pi}(x)$ de $\hat{\theta}_i$ sería $\sqrt{V_{ii}^{\pi}(x)}$, en donde $V_{ii}^{\pi}(x)$ es el i-ésimo elemento de $V^{\pi}(x)$. De manera similar al caso univariante, se tiene la varianza a posteriori del estimador como:

$$\begin{aligned} V_{\hat{\theta}}^{\pi} &= E^{\pi(\theta|x)}\left[\left(\theta - \hat{\theta}\right)\left(\theta - \hat{\theta}\right)^{t}\right] \\ &= V^{\pi}(x) + \left(\mu^{\pi}(x) - \hat{\theta}\right)\left(\mu^{\pi}(x) - \hat{\theta}\right)' \end{aligned} \tag{2.24}$$

Observación: Por abuso del lenguaje se ha notado al parámetro θ de igual manera que su variable aleatoria asociada, por lo que es importante diferenciar entre ambos términos; en el desarrollo anterior se utiliza la variable aleatoria.

2.2.5.3 Conjuntos creíbles

Dentro de la estadística bayesiana la estimación a través de intervalos de confianza se denomina *conjunto creíble.*

***Definición*:** Un conjunto creíble para θ al nivel $100 * (1 - \alpha)\%$, es un conjunto $C \subset \Theta$ tal que:

$$1 - \alpha \leq P(C|x) = \begin{cases} \sum_{\theta \in C} \pi(\theta|x) & caso\ discreto \\ \int_C \pi(\theta|x)dx & caso\ continuo \end{cases} \tag{2.25}$$

Se puede decir que es la probabilidad, subjetivamente hablando (distribución a posteriori), que θ tiene que estar en C. El problema radica en minimizar el tamaño del conjunto creíble; para esto, se puede considerar solamente los puntos con la densidad a posteriori más grande (los valores "más probables").

***Definición*:** El conjunto de máxima densidad a posteriori al $100 * (1 - \alpha)\%$, es el conjunto $C \subset \Theta$, tal que:

$$C = \{\theta \in \Theta: \pi(\theta|x) \geq k(\alpha)\} \tag{2.26}$$

donde $k(\alpha)$ es la constante más grande tal que:

$$P(C|x) \geq 1 - \alpha \tag{2.27}$$

2.2.5.4 Pruebas de hipótesis

Una hipótesis estadística es una proposición o supuesto sobre los parámetros de una o más poblaciones. De manera usual, se tienen dos tipos de hipótesis:

- La hipótesis nula, representada por H_0, es la afirmación sobre una o más características de la población que al inicio se supone cierta ("creencia a priori").
- La hipótesis alternativa, representada por H_1, es la afirmación contradictoria a H_0 y es la hipótesis del investigador.

En la estadística clásica para decidir entre dos hipótesis, se plantea lo siguiente:

$$\begin{cases} H_0: \theta \in \Theta_0 \\ H_1: \theta \in \Theta_1 \end{cases} \tag{2.28}$$

donde, $\Theta_0 \cup \Theta_1 = \Theta$ y $\Theta_0 \cap \Theta_1 = \emptyset$. El procedimiento de decidir entre ambas hipótesis se basa en las probabilidades del *error de primera especie (error tipo I)* y el *error de segunda especie (error tipo II)*, las mismas que representan la posibilidad de rechazar la hipótesis nula dado que es verdadera o de no rechazar la hipótesis nula cuando es falsa, respectivamente.

Desde el punto bayesiano, la decisión entre H_0 y H_1es más simple que en el sentido clásico; pues, solamente se calculan las probabilidades a posteriori de cada una de las hipótesis:

$$\begin{aligned} \alpha_0 &= P(\Theta_0|x) \\ \alpha_1 &= P(\Theta_1|x) \end{aligned} \tag{2.29}$$

Donde, α_0 y α_1 son las probabilidades obtenidas de los datos observados y/o las opiniones de expertos a priori. A las probabilidades a priori de Θ_0 y Θ_1 se les denota por π_0 y π_1, respectivamente.

***Definición*:**

- Se llama razón a priori de H_0 contra H_1 al cociente:

$$\frac{\pi_0}{\pi_1} \tag{2.30}$$

- Se llama razón a posteriori de H_0 contra H_1 al cociente:

$$\frac{\alpha_0}{\alpha_1} \tag{2.31}$$

La interpretación de los cocientes es inmediata. Si la razón a priori es cercana a 1, significa que H_0 y H_1 tienen "casi" la misma probabilidad de ocurrir. Si la razón a priori es mayor que 1, significa que H_0 es más probable que ocurra con respecto a H_1 y si es menor que 1, H_0 es menos probable de ocurrir que H_1.

***Definición (Factor de Bayes)*:**

Se define el factor de Bayes en favor de Θ_0, a la cantidad:

$$B = \frac{\frac{\alpha_0}{\alpha_1}}{\frac{\pi_0}{\pi_1}} = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} \tag{2.32}$$

Nótese que *B* proporciona una medida de la forma en que los datos aumentan o disminuyen las razones de probabilidades de H_0 respecto a H_1. Es decir,

- Si $B > 1$, H_0 es relativamente más probable que H_1.
- Si $B < 1$, la probabilidad relativa de H_1 ha aumentado.

Pruebas de hipótesis de una cola (unilateral)

Sea $\Theta \subseteq \mathbb{R}$, se pueden estudiar los siguientes casos:

$$\begin{matrix} H_0: \theta \leq \theta_0 \\ H_1: \theta > \theta_1 \end{matrix} \quad o \quad \begin{matrix} H_0: \theta \geq \theta_0 \\ H_1: \theta < \theta_1 \end{matrix} \tag{2.33}$$

En estos casos se puede utilizar el $Valor - p$ de manera análoga que en el enfoque clásico.

Contraste de hipótesis nula puntual

En este caso, se considera la siguiente prueba de hipótesis:

$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases} \tag{2.34}$$

Desde el punto de vista bayesiano esta prueba contiene nuevas características con respecto al punto de vista clásico. Cabe recalcar que $\theta = \theta_0$ se cumple, generalmente, de manera asintótica. Por tanto, es más razonable considerar un intervalo de la siguiente forma:

$$\theta \in \Theta_0 = (\theta_0 - b, \theta_0 + b) \tag{2.35}$$

donde, *b* es una constante mayor que 0 que se elige de tal manera que todo $\theta \in \Theta_0$ es "*indistinguible"* de θ_0. Pero, b debe ser lo más pequeño posible; en caso contrario, el contraste tendrá resultados dudosos. Luego, se tiene la prueba de hipótesis:

$$\begin{cases} H_{0'}: \theta \in (\theta_0 - b, \theta_0 + b) \\ H_{1'}: \theta \notin (\theta_0 - b, \theta_0 + b) \end{cases} \tag{2.36}$$

La pregunta que surge es: ¿Cuándo es adecuado aproximar H_0 con $H_{0'}$? Desde la estadística bayesiana, la única respuesta es: la aproximación es adecuada si las probabilidades *a posteriori* de H_0 y $H_{0'}$ son casi iguales. Una condición necesaria es que la función de verosimilitud observada sea aproximadamente constante en $(\theta_0 - b, \theta_0 + b)$.

Para poder realizar el contraste de hipótesis bayesiano para una hipótesis nula puntual $H_0: \theta = \theta_0$, no se utiliza una densidad *a priori* continua, ya que en cualquier caso θ_0 tendrá una probabilidad a priori igual a 0. Es recomendable utilizar una distribución mixta que asigne una probabilidad π_0 al punto θ_0 y $1 - \pi_0$ en el resto de puntos; es decir, la densidad $\pi_1 g_1(\theta)$, con una densidad propia g_1, donde $\pi_1 = 1 - \pi_0$.

Considere una muestra aleatoria simple $X_1, \dots, X_n$ de una variable aleatoria X; la densidad marginal de X es:

$$m = \pi_0 f(x|\theta_0) + (1 - \pi_0) m_1(x) \tag{2.37}$$

donde,

$$m_1 = \int_{\theta \neq \theta_0} f(x|\theta) g_1(\theta) d\theta \tag{2.38}$$

es la densidad bajo H_1. Luego, la probabilidad a posteriori de $\theta = \theta_0$ es:

$$\begin{aligned} \alpha_0 &= P(\theta_0|x) \\ &= \frac{f(x|\theta_0)\pi_0}{m(x)} \\ &= \frac{f(x|\theta_0)\pi_0}{f(x|\theta_0)\pi_0 + \pi_1 m_1(x)} \end{aligned} \tag{2.39}$$

Luego,

$$\frac{\alpha_0}{\alpha_1} = \frac{P(\theta_0|x)}{1 - \pi(\theta_0|x)} = \frac{f(x|\theta_0)\pi_0}{\pi_1 m_1(x)} \tag{2.40}$$

Entonces, el factor de Bayes para la prueba de hipótesis H_0 contra H_1 es:

$$B = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{f(x|\theta_0)}{m_1(x)} \tag{2.41}$$

2.2.5.5 Predicciones

En estadística se persigue un fin común: poder determinar con cierto grado de precisión lo que va a ocurrir con valores futuros que pueden entrar en un conjunto de datos.

Una de las situaciones más comunes es cuando se tiene una variable aleatoria Y, con una distribución de densidad $g(y|\theta)$ (θ desconocida), y se desea predecir un valor de Y cuando se tienen datos de una variable X, derivados de una densidad $h(x|\theta)$. Por ejemplo, si los datos provienen de una regresión, Y sería la variable dependiente de la cual se desea obtener una predicción de una respuesta futura.

Si X e Y son independientes y g es una densidad (en el caso que no lo sean se usaría $g(y|\theta, x)$. Entonces, la idea consiste en determinar la distribución a posteriori de Y dado X.

Definición*:** Se define a la ***densidad predictiva de Y dado $X = x$, cuando la distribución *a priori* para θ es π y se define por:

$$P(y|x) = \int_{\Theta} g(y|\theta)\pi(\theta|x)d\theta \tag{2.42}$$

2.3 ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

En muchos casos es necesario realizar un análisis sobre la relación que existen entre dos o más variables. Este procedimiento se adapta a un amplio rango de fenómenos, desde económicos hasta aspectos humanos.

Como es conocido, este análisis pretende explorar y cuantificar la relación entre una variable dependiente (o criterio), denotada generalmente por Y, y variables independientes (o

predictoras), denotadas por $X_1, \dots, X_n$; así como, desarrollar una ecuación lineal con fines predictivos. Dicha ecuación se escribe de la siguiente manera:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \tag{2.43}$$

De donde se derivan todos los conceptos y aplicaciones conocidos de este análisis. Desde el punto de vista bayesiano, el análisis de regresión lineal múltiple es muy similar al análisis común salvo que en el análisis bayesiano se incluye una distribución a priori como distribución de los parámetros.

2.3.1 DISTRIBUCIÓN A POSTERIORI A PARTIR DE UNA DISTRIBUCIÓN A PRIORI NO INFORMATIVA

Consideremos el siguiente modelo, escrito en su forma matricial:

$$Y = X\beta + \varepsilon \tag{2.44}$$

donde, ε es el vector de los errores con una distribución $N(0, \sigma^2 I)$. Luego, la función de verosimilitud se escribe de la siguiente manera:

$$L(\beta, \sigma^2 | Y) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right] \tag{2.45}$$

como es conocido,

$$\hat{\beta}^{\pi} = (X'X)^{-1}X'Y \qquad y \qquad \hat{Y} = X\hat{\beta}^{\pi} \tag{2.46}$$

Notación: Se notará por $\hat{\beta}^{\pi}$ al estimador de MCO bayesiano, para no confundirlo con el estimador clásico.

Entonces, el producto $(Y - X\beta)'(Y - X\beta)$, se puede calcular como:

$$(Y - X\beta)'(Y - X\beta) = (Y - X\beta)'(Y - X\beta) - 2Y'\hat{Y} + 2Y'\hat{Y}$$

$$= \left(Y - X\hat{\beta}^{\pi}\right)'\left(Y - X\hat{\beta}^{\pi}\right) + \left(\beta - \hat{\beta}^{\pi}\right)'X'X\left(\beta - \hat{\beta}^{\pi}\right)$$

$$= (n-r)\left(\frac{1}{n-r}\right)\left(Y - X\hat{\beta}^{\pi}\right)'\left(Y - X\hat{\beta}^{\pi}\right) + \left(\beta - \hat{\beta}^{\pi}\right)'X'X\left(\beta - \hat{\beta}^{\pi}\right)$$

$$= (n-r)S^2 + \left(\beta - \hat{\beta}^{\pi}\right)'X'X\left(\beta - \hat{\beta}^{\pi}\right) \quad (2.47)$$

Con

$$S^2 = \frac{\left(Y - X\hat{\beta}^{\pi}\right)'\left(Y - X\hat{\beta}^{\pi}\right)}{n-r}$$

Entonces, la función de verosimilitud se reescribe como:

$$L(\beta, \sigma^2|Y) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\left\{(n-r)S^2 + \left(\beta - \hat{\beta}^{\pi}\right)'X'X\left(\beta - \hat{\beta}^{\pi}\right)\right\}\right] \quad (2.48)$$

Dado que $\hat{\beta}^{\pi}$ es un parámetro de posición y σ^2 es un parámetro de escala, las distribuciones a priori no informativas son (ya que se puede considerar una distribución uniforme):

$$\pi(\beta) \propto 1 \qquad y \qquad \pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad (2.49)$$

Si además, se asume independencia entre $\hat{\beta}^{\pi}$ y σ^2,

$$\pi(\beta, \sigma^2) = \pi(\beta)\pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad (2.50)$$

Luego, utilizando el teorema de Bayes se tiene que la distribución a posteriori $\pi(\beta, \sigma^2|Y)$ es,

$$\pi(\beta, \sigma^2|Y) \propto \pi(\beta, \sigma^2)\, L(\beta, \sigma^2|Y) \quad (2.51)$$

Luego, reemplazando, se tiene:

$$\pi(\beta, \sigma^2|Y) = \frac{1}{\sigma^{n+2}}\exp\left\{-\frac{1}{2\sigma^2}\left[(n-r)S^2 + \left(\beta - \hat{\beta}^{\pi}\right)'X'X\left(\beta - \hat{\beta}^{\pi}\right)\right]\right\} \quad (2.52)$$

Dada la multiplicabilidad de la probabilidad, se puede expresar como:

$$\pi(\beta, \sigma^2|Y) \propto \pi(\sigma^2|Y)\pi(\beta|Y, \sigma^2) \quad (2.53)$$

Ahora, es necesario calcular la distribución marginal de $\pi(\sigma^2|Y)$:

$$\pi(\sigma^2|Y) = \int_{-\infty}^{\infty} \pi(\beta, \sigma^2|Y) d\beta$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-r)S^2 + \left(\beta - \hat{\beta}^{\pi}\right)' X' X \left(\beta - \hat{\beta}^{\pi}\right)\right]\right\} d\beta$$

$$= \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2}(n-r)S^2\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\beta - \hat{\beta}^{\pi}\right)' X' X \left(\beta - \hat{\beta}^{\pi}\right)\right]\right\} d\beta$$

$$= \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2}(n-r)S^2\right\} (2\pi)^r \left|\frac{(X'X)^{-1}}{\sigma^2}\right|^{1/2}$$

$$\propto \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2}(n-r)S^2\right\} (2\pi)^r \frac{1}{[(\sigma^2)^{2r}]^{-\frac{1}{2}}}$$

$$\propto \frac{1}{\sigma^{\frac{n-r}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}(n-r)S^2\right\} \tag{2.54}$$

Esta distribución se puede aproximar por una Normal –Gamma-Inversa (la definición se puede ver en el anexo A), que se asumirá como la distribución marginal a posteriori para σ^2, por tanto:

$$\pi(\sigma^2|Y) = GInv\left(\frac{n-r}{2}, \frac{(n-r)S^2}{2}\right) \tag{2.55}$$

Por otro lado, la distribución a posteriori para β, $\pi(\beta|Y, \sigma^2)$, es:

$$\pi(\beta|Y, \sigma^2) = \frac{\pi(\beta, \sigma^2|Y)}{\pi(\sigma^2|Y)}$$

$$\propto \frac{\frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-r)S^2 + \left(\beta - \hat{\beta}^{\pi}\right)' X' X \left(\beta - \hat{\beta}^{\pi}\right)\right]\right\}}{\frac{1}{\sigma^{\frac{n-r}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}(n-r)S^2\right\}}$$

$$= \frac{1}{\sigma^r} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\beta - \hat{\beta}^{\pi}\right)' X' X \left(\beta - \hat{\beta}^{\pi}\right)\right]\right\} \tag{2.56}$$

Entonces, la distribución a posteriori de β, se pude aproximar a una normal multivariante. Ahora, si se integra la expresión de $\pi(\beta, \sigma^2|Y)$ con respecto a σ^2, se obtiene la distribución marginal de β:

$$\pi(\beta|Y) = \int_0^{\infty} \frac{1}{\sigma^{n+2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-r)S^2 + \left(\beta - \hat{\beta}^{\pi}\right)' X' X \left(\beta - \hat{\beta}^{\pi}\right)\right]\right\} d\sigma^2$$

$$= \int_0^{\infty} \sigma^{-\left(\frac{n}{2}+1\right)} \exp\left\{-\sigma^{-2}\left[\frac{(n-r)S^2 + \left(\beta - \hat{\beta}^{\pi}\right)' X' X \left(\beta - \hat{\beta}^{\pi}\right)}{2}\right]\right\} d\sigma^2$$

$$= \left(\frac{(n-r)S^2 + \left(\beta - \hat{\beta}^{\pi}\right)' X' X \left(\beta - \hat{\beta}^{\pi}\right)}{2}\right)^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)$$

$$= \Gamma\left(\frac{(n-r)+r}{2}\right) 2^{\frac{n}{2}} \left\{(n-r)S^2 \left[\frac{(n-r)S^2 + \left(\beta - \hat{\beta}^{\pi}\right)' X' X \left(\beta - \hat{\beta}^{\pi}\right)}{(n-r)S^2}\right]\right\}^{-\frac{1}{2}[(n-r)+r]}$$

$$\propto \Gamma\left(\frac{(n-r)+r}{2}\right) S^{-r} (n-r)^{-\frac{r}{2}} \left\{1 + \frac{\left(\beta - \hat{\beta}^{\pi}\right)' X' X \left(\beta - \hat{\beta}^{\pi}\right)}{(n-r)S^2}\right\}^{-\frac{1}{2}[(n-r)+r]} \tag{2.57}$$

De esta manera se puede ver que la distribución obtenida corresponde a una t-Student multivariada (la definición se puede ver en el anexo A); entonces, se puede concluir que cada $\beta_i, i = 1, \dots, r$ tiene una distribución t-Student univariada con $(n-r)$ grados de libertad.

2.3.2 DISTRIBUCIÓN A POSTERIORI CON INFORMACIÓN A PRIORI CONJUGADA

En este caso se considera que los parámetros de interés en el análisis (β y σ^2) tienen una distribución a priori conocida. Las distribuciones a priori conocidas corresponden a las distribuciones a posteriori encontradas en la sección anterior, ya que la distribución normal-

gamma-inversa es conjugada de la verosimilitud con respecto al vector Y; esto se puede ver a continuación:

$$NGInv(\beta,\sigma^2|Y) \propto \exp\left\{\frac{1}{2\sigma^2}\left(\beta-\hat{\beta}^{\pi}\right)'(X'X)\left(\beta-\hat{\beta}^{\pi}\right)\right\}\left(\frac{1}{\sigma^2}\right)^{\frac{n-r}{2}+1}\exp\left(-\frac{(n-r)S^2}{2\sigma^2}\right)$$

$$\propto \exp\left\{\frac{1}{2\sigma^2}[(n-r)S^2+\left(\beta-\hat{\beta}^{\pi}\right)'(X'X)\left(\beta-\hat{\beta}^{\pi}\right)]\right\}$$

$$\propto L(\beta,\sigma^2|Y) \qquad (2.58)$$

Las distribuciones a posteriori $\pi(\sigma^2|Y)$ y $\pi(\beta|\sigma^2,Y)$ calculadas en la muestra original. Ahora se consideran las distribuciones a priori conjugadas.

$$\pi(\sigma^2) = GInv\left(\frac{n_1-r}{2},\frac{(n_1-r)S_1^2}{2}\right) \qquad y \qquad \pi(\beta|\sigma^2) = N\left(\hat{\beta}^{\pi}{}_1,\sigma^2(X'X)^{-1}\right) \qquad (2.59)$$

donde n_1 es el tamaño de la muestra original, utilizando los mismos estimadores calculados anteriormente. Adicionalmente, supóngase que se toma una segunda muestra de tamaño n_2 de donde se obtiene la siguiente función de verosimilitud:

$$L(\beta,\sigma^2|Y_2) = \frac{1}{(2\pi)^{\frac{n_2}{2}}\sigma^{n_2+r}}\exp\left[-\frac{1}{2\sigma^2}(Y_2-X_2\beta)'(Y_2-X_2\beta)\right] \qquad (2.60)$$

A partir de los mínimos cuadrados ordinarios, el estimador de β tiene la siguiente estructura:

$$\hat{\beta}^{\pi} = (X_1'X_1+X_2'X_2)^{-1}(X_1'X_1+X_2'X_2) \qquad (2.61)$$

Utilizando el teorema de Bayes, la distribución a posteriori conjunta de β y σ^2 se describe por:

$$\pi(\beta,\sigma^2|Y_2) \propto l(\beta,\sigma^2|Y_2)\pi(\beta,\sigma^2)$$

$$= \frac{1}{(2\pi)^{\frac{n_2}{2}}\sigma^{n_2+r}}\exp\left[-\frac{1}{2\sigma^2}\left\{(n_2-r)S_2^2+\left(\beta-\hat{\beta}^{\pi}\right)'X_2'X_2\left(\beta-\hat{\beta}^{\pi}\right)\right\}\right]*$$

$$\left(\frac{1}{\sigma^2}\right)^{\frac{n_1}{2}+1}\exp\left[-\frac{1}{2\sigma^2}\left\{(n_1-r)S_1^2+\left(\beta-\hat{\beta}\right)'X_1'X_1\left(\beta-\hat{\beta}\right)\right\}\right] \qquad (2.62)$$

donde,

$$S_2^2 = \frac{\left(Y_2 - X_2\hat{\beta}^{\pi}\right)'\left(Y_2 - X_2\hat{\beta}^{\pi}\right)}{n_2 - r}$$

Luego, por la ley multiplicativa de la probabilidad, se puede escribir:

$$\pi(\beta, \sigma^2|Y_2) = \pi(\sigma^2|Y_2)\pi(\beta|\sigma^2, Y_2) \tag{2.63}$$

De donde, siguiendo el razonamiento utilizado para calcular las distribuciones marginales a posteriori de σ^2 y β con una distribución a priori no informativa, se puede concluir que:

$$\pi(\sigma^2|Y_2) = GInv\left(\frac{n_1 + n_2 - r}{2}, \frac{(n_1 + n_2 - r)(S^*)^2}{2}\right) \tag{2.64}$$

donde,

$$(S^*)^2 = \frac{1}{n_1 + n_2 - r}\left[\left(Y_1 - X_1\hat{\beta}^{\pi}\right)'\left(Y_1 - X_1\hat{\beta}^{\pi}\right) + \left(Y_2 - X_2\hat{\beta}^{\pi}\right)'\left(Y_2 - X_2\hat{\beta}^{\pi}\right)\right] \tag{2.65}$$

De manera análoga al caso de considerar una distribución a priori no informativa, la distribución marginal a posteriori de β es una t-Student multivariante y cada β_i se distribuye como una t-Student univariante con $n_1 + n_2 - r$ grados de libertad. Además,

$$\hat{\beta}^{\pi} = (X_1'X_1 + X_2'X_2)^{-1}(X_1'Y_1 + X_2'Y_1) \tag{2.65}$$

2.3.3 INFERENCIA BAYESIANA PARA σ^2 Y β

Como en el caso del análisis de regresión múltiple común, una vez encontradas las distribuciones de los parámetros es necesario tener la inferencia de cada uno; esto se describe en esta sección.

2.3.3.1 Estimación puntual

Para el cálculo de la estimación del parámetro β se necesita calcular la moda a posteriori; sin embargo, dado que la distribución marginal de β es una t-Student multivariante, la moda a posteriori coincide con la media y la mediana por la simetría, entonces $\hat{\beta}^{\pi}$ de ec(2.65) es el estimador de β.

Para σ^2 se calcula la moda de la distribución gamma-inversa que coincide con el estimador de máxima verosimilitud de $\pi(\sigma^2|Y)$, por lo tanto:

$$\pi(\sigma^2|Y) = \frac{(n-r)S^2}{2\Gamma\left(\frac{n-r}{2}\right)}\sigma^{\frac{n-r}{2}+1}\exp\left(\frac{(n-r)S^2}{2\sigma^2}\right) \tag{2.66}$$

Luego, aplicando el logaritmo natural a la ecuación, derivando con respecto a σ^2 e igualando a 0 y resolviendo, se obtiene:

$$\sigma^2_{moda} = \frac{(n-r)S^2}{n-r+2} \tag{2.67}$$

Sin embargo, la distribución gamma-inversa no es simétrica por lo que la media y la mediana no coinciden con la moda; se elige la media que tenga la menor varianza, esto es:

$$\mu^{\pi} = E^{\pi(\sigma^2|Y)}[\sigma^2] = \frac{\frac{(n-r)S^2}{2}}{\frac{n-r}{2}-1} = \frac{(n-r)S^2}{n-r-2}, \qquad para\ \frac{n-r}{2} > 2 \tag{2.68}$$

el error de estimación para σ^2_{moda} es:

$$V^{\pi}_{\sigma^2_{moda}} = V^{\pi} + (\mu^{\pi} - \sigma^2_{moda})^2$$

$$= \frac{2(n-r)^2S^4}{(n-r-2)^2(n-r-4)} + \left(\frac{(n-r)S^2}{n-r-2} - \frac{(n-r)S^2}{n-r+2}\right)^2$$

$$= \frac{2(n-r)^2 S^4}{(n-r-2)^2(n-r-4)} + \frac{16(n-r)^2 S^4}{((n-r)^2-4)^2} \tag{2.69}$$

2.3.3.2 Conjuntos creíbles de máxima densidad a posteriori $\boldsymbol{\beta}$

Dado que la distribución marginal a posterior de β_i es una t-Student univariante, un conjunto creíble se calcula de la siguiente manera:

$$\left[\hat{\beta}^{\pi}{}_i - t_{\frac{\alpha}{2}}\sqrt{V_{ii}}\ ;\ \hat{\beta}^{\pi}{}_i + t_{\frac{\alpha}{2}}\sqrt{V_{ii}}\right] \tag{2.70}$$

donde, $t_{\frac{\alpha}{2}}$ es el cuantil de orden $\frac{\alpha}{2}$ de la distribución t-Student con $(n-r)$ grados de libertad. Y los V_{ii} es el elemento (i,i) de la matriz: $S^2(X_1'X_1 + X_2'X_2)^{-1}$.

2.4 ANÁLISIS BAYESIANO DE LA REGRESIÓN LOGÍSTICA

Como es conocido en el modelo de regresión logística se considera una variable respuesta que tiene dos categorías de respuesta y sigue una distribución de Bernoulli. Si se tuvieran n observaciones:

$$y_i|\pi_i \sim Be(\pi_i)$$

$$\theta = \pi_i = P\{y_i = 1\} = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}, i = 1, \dots n \tag{2.71}$$

donde, $\beta = (\beta_1, \dots, \beta_r)'$ vector de parámetros desconocidos y $X' = (1, x_{i1}, \dots, x_{ir})'$ es el vector de los predictores. Luego, el complemento de la probabilidad de no ocurrencia de un determinado evento es:

$$1 - \theta = \frac{1}{1+e^{x_i'\beta}} \tag{2.72}$$

2.4.1 LA VEROSIMILITUD

La función de verosimilitud de los parámetros desconocidos β, es:

$$L(\beta|Y,X) = \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + x_i\beta_1}}{1 + e^{\beta_0 + x_i\beta_1}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + y_i\beta_1}} \right)^{1-y_i} \tag{2.73}$$

El estimador de máxima verosimilitud β se obtiene maximizando el logaritmo de la función de verosimilitud. En este caso, se tiene:

$$\hat{\beta}^{\pi} \sim N_{r+1}\left(\beta, I\left(\hat{\beta}^{\pi}\right)^{-1}\right) \tag{2.74}$$

donde, $I\left(\hat{\beta}^{\pi}\right)$ es la información de Fisher. Las regiones creíbles se pueden encontrar a partir de la función de verosimilitud.

2.5 ANÁLISIS BAYESIANO DE COMPONENTES PRINCIPALES

Dentro del análisis de componentes principales se busca reducir la dimensionalidad de un conjunto de datos, de tal manera que los datos queden "bien" representados en términos de mínimos cuadrados. Una de las ventajas del ACP para reducir la dimensión de un conjunto de datos, es que retiene las características del conjunto de datos que contribuyen más a su varianza, tratando de tener el menor número de componentes que logren el objetivo.

Previo a la presentación de los conceptos bayesianos de las componentes principales, es necesario introducir algunas definiciones que serán de ayuda para comprender esta sección.

Si se considera X una matriz de orden nxp (en general, $n > p$), donde en las columnas están contenidas las variables y en las filas los elementos. La idea del ACP es encontrar un espacio de dimensión $r < p$, en el cual los puntos estén mejor representados los puntos de la matriz X a través de los r mayores valores propios de la matriz de covarianzas S. Las componentes principales se calculan mediante:

$$|S - \lambda I| = 0 \tag{2.75}$$

donde,

$\lambda = (\lambda_1, \dots, \lambda_p)'$ es el vector de los valores propios de la matriz S.

p: es el rango de S.

Además, los vectores propios asociados se determinan a través de:

$$(S - \lambda_i I)a_i = 0 \tag{2.76}$$

Se denomina Z a la matriz cuyas columnas son los valores de las p componentes en los n individuos; estas nuevas variables están relacionadas por:

$$Z = XA \tag{2.77}$$

donde, $A'A = I$. A se denomina la matriz de cargas factoriales; una carga factorial se la define como la correlación entre una variable original y un factor, por lo tanto la matriz de cargas factoriales no es más que una matriz que contiene las correlaciones entre las variables originales y los factores. Las componentes principales equivale a aplicar una transformación ortogonal A (matriz de cargas factoriales) a las variables de X para obtener unas nuevas variables Z no correlacionadas entre sí.

Ahora, ya en el contexto bayesiano la distribución *a posteriori* se puede definir como la densidad condicional de las medidas consideradas dentro del ACP tradicional; esto se puede escribir como:

$$P\left(\tilde{Z}, \tilde{\alpha}, \tilde{r} \middle| X\right) = \frac{P\left(X \middle| \tilde{Z}, \tilde{\alpha}, \tilde{r}\right) P\left(\tilde{Z}, \tilde{\alpha}, \tilde{r}\right)}{P(X)} \tag{2.78}$$

donde,

$\tilde{Z}$: Es la matriz de los componentes principales de orden $(n\ x\ p)$.

$\tilde{\alpha}$: Es la matriz ortogonal de las cargas factoriales de orden $(p\ x\ p)$.

$\tilde{r}$: Es el rango verdadero de la matriz de datos X (el número de componentes principales, $\tilde{r} < p$).

X: Matriz de datos.

n: Número de observaciones.

p: Número de variables.

Observación: Cabe recalcar que las componentes principales al ser una combinación lineal de todas las características de los datos, pueden estar contaminadas por "ruido" (al ser estimaciones tienen un margen de error con respecto a los datos reales); por tanto, es necesario aplicar algunas técnicas que en lo posible logren aplacar este efecto (Rezghi, M y Obulkasim, A. 2014).

El primer término en el numerador es la función de verosimilitud, que es la densidad condicional de las variables dadas del modelo de componentes principales libre de ruido; mientras que, el segundo término es la distribución a priori. Por lo tanto, la distribución a posteriori (no normalizada) puede ser escrita como:

$$P(\tilde{Z}, \tilde{\alpha}, \tilde{r}|X) \propto P(X|\tilde{Z}, \tilde{\alpha}, \tilde{r})P(\tilde{Z}, \tilde{\alpha}, \tilde{r}) \tag{2.79}$$

2.5.1 LA FUNCIÓN DE DENSIDAD A PRIORI

La distribución a priori es la densidad conjunta de los componentes principales libres de ruido, las cargas factoriales y el rango verdadero del análisis de componentes principales; por tanto, no es una función sencilla de determinar. Sin embargo, la función de densidad de los componentes principales y de las cargas factoriales dependen del rango del modelo. Luego, la probabilidad a priori se puede escribir como:

$$P(\tilde{Z}, \tilde{\alpha}, \tilde{r}) = P(\tilde{Z}, \tilde{\alpha}|\tilde{r})P(\tilde{r}) \tag{2.80}$$

donde,

$$P(\tilde{r}) = P(\tilde{r} = i) = k_i, \quad \sum_{i=1}^{r} k_i = 1 \tag{2.81}$$

Adicionalemente, la función de densidad conjunta de los componentes principales y las cargas factoriales pueden ser expresadas usando la regla de multiplicación de las probabilidades:

$$P\left(\tilde{Z}, \tilde{\alpha} \middle| \tilde{r}\right) = P\left(\tilde{Z} \middle| \tilde{\alpha}, \tilde{r}\right) P(\tilde{\alpha}|\tilde{r}) \tag{2.82}$$

Luego, la distribución a posteriori no normalizada puede ser escrita como:

$$P\left(\tilde{Z}, \tilde{\alpha}, \tilde{r} \middle| X\right) \propto P\left(X \middle| \tilde{Z}, \tilde{\alpha}, \tilde{r}\right) P\left(\tilde{Z} | \tilde{\alpha}, \tilde{r}\right) P(\tilde{\alpha}|\tilde{r}) P(\tilde{r}) \tag{2.83}$$

CAPÍTULO 3
DISEÑO Y CONSTRUCCIÓN DE UN SISTEMA DE ÍNDICES DE PREVENCIÓN Y PROPAGACIÓN DE INCENDIOS FORESTALES (SIPIF)

En este capítulo se presenta el desarrollo de la metodología utilizada para diseñar y construir un sistema de índices de prevención y propagación de incendios forestales (SIPIF) para el Parque Metropolitano Guangüiltagua de Quito.

3.1 DISEÑO DE UN SIPIF

Aquí se propone un sistema de índices que tenga en cuenta, además de las variables meteorológicas, variables consideradas como de riesgo dentro del contexto de incendios forestales; estás son:

Tabla 3. Variables consideradas dentro del estudio

Variables Meteorológicas	**Variables de combustibilidad vegetal**	**Variables Topográficas**
Temperatura	Cantidad de combustible (índice de biomasa)	Pendiente geográfica
Velocidad del viento	Tamaño y forma	Exposición solana
Dirección del viento	Compactación	Altitud
Humedad relativa	Humedad	
Días consecutivos sin lluvia	Distribución espacial	

Elaboración: Autores

Las variables presentadas en la tabla, se definen a continuación:

3.1.1 VARIABLES METEOROLÓGICAS

Son variables que miden una propiedad o condición de la atmósfera, para definir el estado del clima de un lugar determinado, para un período de tiempo dado.

- **Temperatura:** cuantifica la temperatura del ambiente medida en °C.
- **Velocidad del viento:** mide la velocidad del viento en km/h.
- **Dirección del viento:** indica de dónde proviene el viento, su unidad de medición es en grados *Dextrorsum* (giro en sentido de las manecillas del reloj), donde 0° es el norte verdadero (dirección del Polo Norte en relación con la posición del observador).
- **Humedad relativa:** mide el porcentaje de saturación del aire dado por el cociente entre la presión real del vapor de aire a una temperatura dada y la presión de saturación del vapor de aire a la misma temperatura.
- **Días consecutivos sin lluvia:** se calcula como la suma de días a partir de la última precipitación significativa, en los cuales la precipitación es cero o es menor que un umbral determinado.

3.1.2 VARIABLES DE COMBUSTIBILIDAD VEGETAL

Cuantifican el grado de ignición de la cobertura vegetal del terreno de acuerdo a su tipo, tamaño, concentración, entre otros.

- **Cantidad de combustible (índice de biomasa):** mide la cantidad de combustible que puede generar la biomasa vegetal.
- **Tamaño y forma:** indica el tamaño y la forma de la cobertura vegetal presente.
- **Compactación:** mide la densidad de vegetación en un área determinada.
- **Humedad:** cuantifica la humedad presente en la vegetación.
- **Distribución espacial:** indica cómo se encuentran ubicados los diferentes tipos de vegetación en la zona de estudio.

3.1.3 VARIABLES TOPOGRÁFICAS

Proporcionan información sobre las características del terreno en estudio.

- **Pendiente geográfica:** mide el ángulo que forma el plano horizontal con el plano tangente a la superficie del terreno en un punto.
- **Exposición solana:** mide la cantidad de radiación solar que reciben laderas o vertientes de una montaña.
- **Altitud:** mide la distancia vertical que existe entre un punto de la tierra y el nivel del mar.

Junto a estas se consideran variables como: número de incendios ocurridos durante los últimos años, cantidad de hectáreas afectadas por los incendios, ubicación y fecha de ocurrencia de los incendios y tiempo que tomó aplacar el incendio.

3.1.4 METODOLOGÍA

La metodología a utilizada es de dos tipos:

1. La primera parte del sistema de índices se enfoca en determinar la probabilidad de ocurrencia de un incendio a partir de información histórica de los incendios registrados en el Parque Metropolitano desde el año 2012 al 2014. Para ello se utiliza una regresión tipo logístico con estimación bayesiana; en dicha regresión se toman como variables exógenas la temperatura, la velocidad del viento, la radiación solar, la precipitación, el índice de biomasa y el tipo de vegetación, y como variable endógena la ocurrencia o no de un incendio, para observar la relación entre dichas variables y la ocurrencia o no de un incendio; y así determinar un índice de peligrosidad de incendios forestales en base a las probabilidades obtenidas a través del modelo.

2. La segunda parte del sistema de índices busca determinar la posible afectación (hectáreas de bosque quemadas) debido a la propagación de un incendio, que se denomina índice de propagación forestal. En este caso, se realiza una ponderación de

los factores que afectan a la propagación de incendio (variables meteorológicas, ambientales y topográficas) a través de las técnicas: Análisis de Componentes Principales (ACP) y Análisis de Componentes Principales Categórico (ACPC), según se requiera. Posteriormente, con los factores que se determinan, se realiza una Regresión Lineal Múltiple (RLM) con estimación bayesiana y finalmente, se obtienen las ponderaciones (pesos) de los factores dentro del índice.

3.1.5 Obtención y limitaciones de las variables

Una de las grandes falencias que tiene el Ecuador es no tener bases de datos fiables, en diversas áreas de investigación e interés público. Lastimosamente, en este caso sucede lo mismo. A continuación, se realiza una descripción de los problemas encontrados dentro de las bases de datos consideradas dentro de este proyecto:

1. La primera limitación para la obtención de la información del proyecto se da en las variables meteorológicas, debido a que estas variables se obtienen de un sistema telemétrico (que es transmisión por radio de señales procedentes de todo tipo de sensores); existe pérdida de información por la falta de calibración de los equipos, por fallas mecánicas de los equipo o por falta de mantenimiento a estos[4].

2. En cuanto a las variables que resumen los incendios para el Parque Metropolitano Guangüiltagua, se puede decir que el COE no tiene depurada por completo la base de datos; hay muchos datos faltantes y otros que no corresponden a la realidad; por ejemplo, un incendio forestal que dura aproximadamente 1 hora y no ha quemado ningún área de terreno con cobertura vegetal.

[4] Los datos fueron obtenidos de la Estación Meteorológica Bellavista, de la EPMAPS; los técnicos encargados proporcionaron los motivos de la falta de datos.

Adicionalmente, las únicas bases de datos disponibles son solamente las de los años 2013 y 2014; antes del 2013 la base de datos que posee el COE es todavía más deficiente, por lo que no se pudo trabajar con más años y datos.

3. Para el caso de las variables de combustibilidad vegetal y topográficas, la base de datos que se utilizó es la correspondiente al Mapa de Cobertura Vegetal del DMQ, 2010, que fue proporcionado por el Ministerio del Ambiente.

3.2 CONSTRUCCIÓN DEL SIPIF

En esta sección se presenta a detalle la construcción de los índices descritos en el acápite anterior. Se inicia realizando un análisis descriptivo de las variables a utilizarse.

3.2.1 TEMPORALIDAD

Luego de realizar un análisis de los incendios registrados por el COE, se llegó a la conclusión que el 100% de los incendios que se han producido en el DMQ, se llevan a cabo de junio a septiembre en ambos años; por tanto, en este proyecto se utilizan los datos que están dentro de este período de tiempo para todas las variables, excepto las de cobertura vegetal y topográficas.

3.2.2 ANÁLISIS DESCRIPTIVO DE LAS VARIABLES

En esta sección se realiza un análisis descriptivo de cada una de las variables para poder determinar la información que puede contribuir al desarrollo de este proyecto. Para iniciar, se consideran las variables meteorológicas:

Tabla 4. Estadísticas descriptivas de las variables meteorológicas

	Humedad Relativa (%)	**Velocidad del Viento** $(\mathbf{km}/\mathbf{h})$	**Temperatura (°C)**	**Días consecutivos sin lluvia**
Mínimo	15,66	0,45	8,06	0,00
Máximo	92,44	7,93	25,70	79,00
Promedio	60,96	4,19	14,75	28,18
Número de datos	5.438	5.438	5.438	5.438

Fuente: Estación Meteorológica Bellavista, EPMAPS.
Elaboración: Autores

Como se puede ver en la tabla anterior, se tiene un total de 5.438 datos; cabe aclarar que se han considerado los datos cada hora desde el 1 de junio de 2013 a las 0:00 horas hasta el 30 de septiembre de 2014 a las 23:00 y deberían existir 5.856 datos; sin embargo, por la pérdida de información descrita en la sección anterior, se pierde un 7% de los datos.

Por otro lado, se puede observar también que la temperatura promedio es de 14,75 °C, el promedio de velocidad de viento es de 4,19 km/h y la humedad relativa promedio es del 61%. Es importante observar que el número de días secos en los meses de junio a septiembre llegaron a incluso 79, con un promedio de 28 días sin recibir lluvia.

De la misma manera se realiza un análisis descriptivo de los incendios y de las variables inmersas en ellos, cuyos resultados se presentan a continuación:

Tabla 5. Estadísticas descriptivas de los incendios

	Incendios	Tiempo de incendio (h:min:seg)	Área quemada (m^2)
Total	23	-	-
Mínimo	-	0:06:00	0
Máximo	-	3:23:00	20.000
Promedio	-	1:25:46	1.017

Fuente: COE – Q
Elaboración: Autores

En la tabla anterior, se muestran estadísticos descriptivos del tiempo trascurrido y área quemada por los incendios que ocurrieron en el periodo antes mencionado. Se registraron 23 incendios dentro del Parque Metropolitano Guangüiltagua. Los incendios tuvieron una duración mínima de seis minutos y una máxima de tres horas y veinte y tres minutos, con un promedio de una hora y veinte y cinco minutos; quemando un área promedio de 1.017 metros cuadrados.

3.2.3 DEPURACIÓN DE LA BASE DE DATOS DE INCENDIOS

Una vez identificada la temporalidad de los datos y las limitaciones dadas por los proveedores de las bases de datos, fue necesario realizar una depuración de las mismas, principalmente de incendios.

Se encontraron incendios con duración de una hora y que tenían cero metros cuadrados quemados; por lo tanto, se analizó cada uno de ellos y se logró determinar que muchos correspondían a quemas controladas (de basura o vegetación muerta). Finalmente, se determinó que existían 14 incendios forestales dentro del Parque Metropolitano Guangüiltagua, de los que se tenía información completa y coherente.

Por otro lado, la base correspondiente a las variables meteorológicas no fue menos problemática, debido a pérdida de información en la recolección de los datos por parte de la Estación Meteorológica Bellavista. Es así como, en este caso se detectaron señales mínimas constantes en cada una de las variables en algunos períodos de tiempo, debido a que los sensores dejaron de funcionar o fueron recalibrados. Estos puntos se eliminaron de la base de datos del proyecto.

Finalmente, se obtuvo una base de datos que contenía 5.429 datos, de los cuales 14 correspondían a incendios.

3.2.4 JUSTIFICACIÓN DE LAS TÉCNICAS UTILIZADAS EN EL PROYECTO

Una de las preguntas básicas en el estudio consiste en contestar por qué se utilizan métodos bayesianos y no los tradicionales. Esto se justifica debido a la poca cantidad de datos sobre los incendios forestales que se tienen en la base; la distribución a priori ayuda a la estimación cuando se dispone de poca cantidad de datos.[5] Por otro lado, considerar los parámetros de las ecuaciones como variables aleatorias (y no estáticamente) contribuye a una mejor estimación de las probabilidades de ocurrencia de un incendio y la estimación de la propagación del mismo.

3.3 ESTIMACIÓN DE MODELOS ESTADÍSTICOS

Una vez que se han depurado las bases de datos y determinado los datos válidos a utilizarse, se inicia con el modelamiento de los índices. En este proyecto se utiliza el paquete estadístico R versión 3.2.2.

3.3.1 ÍNDICE DE PREVENCIÓN DE INCENDIOS

Este índice pronostica la probabilidad de que un incendio ocurra. Para ello se utiliza la estimación de un modelo de regresión logística bayesiana (en este caso, se considera como estimador bayesiano a la media de la distribución a posteriori). Se inicia considerando todas las variables para el modelo; estas son: Temperatura, Humedad Relativa, Velocidad del Viento, los días secos consecutivos, la cobertura vegetal ingresada como variables *dummy* por cada tipo de vegetación y una variable respuesta dicotómica que toma el valor 1 cuando ha ocurrido un incendio y 0 en caso contrario. Cabe recalcar que 0 corresponde a la hora en la que no ha ocurrido un incendio y se considera como cobertura vegetal una categoría denominada "*Todo"* que abarca a todos los tipos de cobertura vegetal. Como se puede prever existen una gran cantidad de "0" en contraste con catorce valores "1" en la variable

[5] Berry, DA. 1996. Statistics: A Bayesian perspective. Belmont, California, Duxbury Press

dependiente. Así, tomando como distribución a *priori* una distribución normal multivariada (en adelante a menos que se indique lo contrario, siempre se considera de la misma manera la distribución *a priori*) se obtiene:

Tabla 6. Modelo 1. Regresión logística bayesiana para el índice de prevención de incendios considerando todas las variables

Coefficients:				
	Estimate	Std. Error	z.value	Pr(>|z|)
(Intercept)	-6.363479	4.844876	-1.313	0.18903
Temp	0.064800	0.166724	0.389	0.69752
HR1	-0.043998	0.040729	-1.080	0.28003
Vel.viento	0.135539	0.182943	0.741	0.45877
Días.secos	-0.001169	0.018730	-0.062	0.95022
Arb. Mont. andes del norte`	4.921542	1.767198	2.785	0.00535 **
Arbustal secos interandinos`	7.293726	1.116057	6.535	6.35e-11 ***
Eucalipto adulto `	8.279952	1.046009	7.916	2.46e-15 ***
Pasto cultivado`	4.473949	1.704935	2.624	0.00869 **
Pinos y cipres`	4.857846	1.756782	2.765	0.00569 **
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 194.857 on 5428 degrees of freedom
Residual deviance: 24.345 on 5419 degrees of freedom AIC: 44.345

Elaboración: Autores

En este modelo se puede ver que las variables *dummy* referentes a la cobertura vegetal son significativas; sin embargo, todas las demás variables son no significativas. Dado que el modelo así expresado no utilizaría las variables meteorológicas, se decide realizar un análisis bivariado de las variables para identificar posibles combinaciones lineales entre las variables. Así, se logró determinar que, por la forma de cálculo, la Humedad Relativa se convierte en función lineal de la temperatura; por tanto, se la sacó del modelo teniendo como resultado el modelo descrito a continuación:

Tabla 7. Modelo 2. Modelo de regresión logística bayesiana sin considerar la variable Humedad Relativa

Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-11.225.191	2.229.693	-5.034	4.79e-07	***
Temp	0.212598	0.104611	2.032	0.04213	*
Vel.viento	0.142731	0.182573	0.782	0.43435	
Días.secos	0.003383	0.018008	0.188	0.85100	
datosd$`Arbustal montano de los andes del norte`	4.978.220	1.780.467	2.796	0.00517	**
datosd$`Arbustal secos interandinos`	7.351.899	1.100.831	6.678	2.41e-11	***
datosd$`Eucalipto adulto`	8.341.432	1.045.064	7.982	1.44e-15	***
datosd$`Pasto cultivado`	4.640.714	1.724.862	2.690	0.00713	**
datosd$`Pinos y cipres`	4.950.440	1.776.345	2.787	0.00532	**
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 194.857 on 5428 degrees of freedom
Residual deviance: 24.848 on 5420 degrees of freedom AIC:42.848

Elaboración: Autores

Se inicia un proceso iterativo de eliminación de las variables no significativas, se elimina la variable menos significativa[6] del modelo, los Días secos en este caso; sin embargo, siguen existiendo variables que no son significativas. Luego, se elimina la variable velocidad del viento y se obtiene el siguiente modelo:

[6] Se considera que una variable menos significativa a aquella variable que tenga el p-valor más alto.

Tabla 6. Modelo 3. Modelo de regresión logística bayesiana (todas las variables significativas)

```
Coefficients:
                                Estimate   Std.Error  z.value  Pr(>|z|)
(Intercept)                    -10.4066     1.8975    5.484   4.15e-08 ***
Temp                             0.2091     0.1033    2.025   0.04290 *
Arb. Mont. andes del norte`      5.0628     1.7976    2.816   0.00486 **
Arbustal secos interandinos`     7.4030     1.1119    6.658   2.78e-11 ***
Eucalipto adulto`                8.1306     0.9996    8.134   4.17e-16 ***
Pasto cultivado`                 4.8207     1.7463    2.761   0.00577 **
Pinos y cipres`                  5.0381     1.7913    2.813   0.00491 **
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 194.857  on 5428  degrees of freedom
Residual deviance:  25.469  on 5422  degrees of freedom   AIC: 39.469
Number of Fisher Scoring iterations: 12
```

Elaboración: Autores

Se puede observar que solamente quedó la variable temperatura y las variables correspondientes a la cobertura vegetal dentro del modelo; esto quiere decir que el modelo es muy pobre para predecir la probabilidad de ocurrencia de incendio; por ejemplo, considerando un valor de 25 °C (temperatura máxima registrada en los años de análisis) y Eucalipto adulto (tiene el mayor valor del parámetro dentro del modelo con respecto a las demás categorías de cobertura vegetal), se tendría una probabilidad de 0% de que ocurra un incendio, lo que se debe a la falta de datos.

3.3.1.1 Índice de prevención considerando datos de todo el DMQ

Con el fin de mostrar que el procedimiento realizado en este proyecto es pertinente y válido con la condición de que existan un número mayor de datos, se realiza una estimación considerando todos los incendios forestales del DMQ registrados en el COE – Q; pero, considerando como condiciones climatológicas las obtenidas de la Estación Meteorológica

Bellavista que es la información que tenemos disponible. Esto es un tanto arbitrario debido a los microclimas que están presentes en el DMQ.

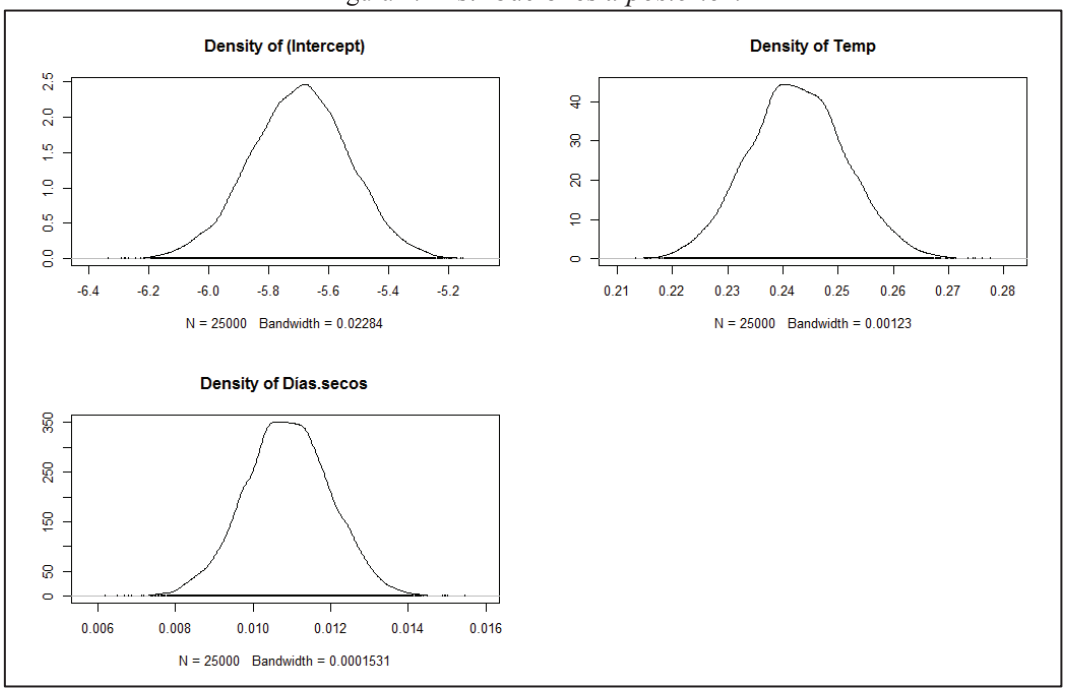
El proceso de modelación es similar al realizado en la sección anterior, con la ventaja de que en este caso el número de incendios permite que el modelo se lo pueda explicar a través de variables meteorológicas y, desafortunadamente, no forman parte del modelo las variables correspondientes a la cobertura vegetal. A continuación, se presenta un modelo general; es decir, un modelo de predicción de probabilidad de inicio de un incendio sin considerar la cobertura vegetal y utilizando como distribución a *priori* una distribución normal multivariada:

Tabla 7. Modelo General del índice de prevención de incendios, con las distribuciones a posteriori de los residuos

Coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.687740	0.163143	-34.864	<2e-16 ***
Temp	0.242061	0.008771	27.599	<2e-16 ***
Días s_lluvia	0.010918	0.001096	9.959	<2e-16 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
(Dispersion parameter for binomial family taken to be 1) Null deviance: 7198.2 on 6813 degrees of freedom Residual deviance: 6161.4 on 6811 degrees of freedom AIC: 6167.4				
Number of Fisher Scoring iterations: 5				

Elaboración: Autores

Figura 2. Distribuciones *a posteriori*



Elaboración: Autores

En la tabla anterior, se puede observar que las variables que ayudan a pronosticar la probabilidad del inicio de un incendio son: temperatura y días secos; además en la figura anterior se observan las distribuciones a posteriori obtenidas, que tal como se podía esperar son normales.

Dentro de la estimación es importante el considerar algunos conceptos que son necesarios para determinar la validez estadística del modelo; por ejemplo, el estimador utilizado y la forma de estimación, los conjuntos creíbles y las pruebas de hipótesis correspondientes a cada uno de los parámetros estimados. En este caso se tiene:

- Los coeficientes estimados corresponden a la media de la distribución a posteriori (dentro del marco de la teoría de la decisión, se puede demostrar que es el mejor estimador del parámetro bayesiano). La estimación se realiza a través de máxima verosimilitud.
- Los conjuntos creíbles son similares a los intervalos de confianza de la estadística clásica; así por ejemplo, para el creciente asociado al intercepto, se tiene:

$$CC = \left[\hat{\beta}_0^{\pi} - t_{1-\frac{\alpha}{2}}\sqrt{V_{ii}}\ ;\ \hat{\beta}_0^{\pi} + t_{1-\frac{\alpha}{2}}\sqrt{V_{ii}}\right]$$
$$= [-5{,}688 - (34.864)(0{,}163)\ ; -5{,}688 + (34.864)(0{,}163)]$$
$$= [-11{,}371; -0{,}005]$$

Para los demás estimadores se procede de manera similar.

- Finalmente, para la validez del modelo se procede con las pruebas de hipótesis correspondientes; por ejemplo, en el caso del intercepto, se tiene:

$$\begin{cases} H_0\colon \hat{\beta}_0^{\pi} = 0 \\ H_1\colon \hat{\beta}_0^{\pi} \neq 0 \end{cases}$$

En este caso, se utiliza el valor-p como en el enfoque clásico. Es decir, dado que el valor-p asociado ($2e^{-16}$) es menor que el nivel de confianza ($\alpha = 0{,}05$), se rechaza la hipótesis nula y, por lo tanto, $\hat{\beta}_0^{\pi}$ es significativo (estadísticamente diferente de cero).

Luego, considerando condiciones promedio y extremo de las variables del modelo final, se obtiene lo siguiente:

1. Temperatura mínima de 8,06 °C y 1 día seco, se obtiene una probabilidad de 2,4% de que un incendio se llegara a producir.

2. Temperatura promedio de 14,75 °C y 39 días secos, se obtiene una probabilidad de 15,56% de que un incendio se produzca.

3. Temperatura máxima de 25,70 °C y 104 días secos, se obtiene una probabilidad del 84,14% de que un incendio se produzca.

Como se puede observar en los ejemplos anteriores, estas probabilidades no son ajenas a los fenómenos que se producen en Quito, específicamente en el Parque Metropolitano Guangüiltahua; por lo que, pueden servir de punto de partida para una construcción más elaborada y con mejores bases de datos para obtener una mejor predicción.

3.3.2 ÍNDICE DE PROPAGACIÓN DE INCENDIOS

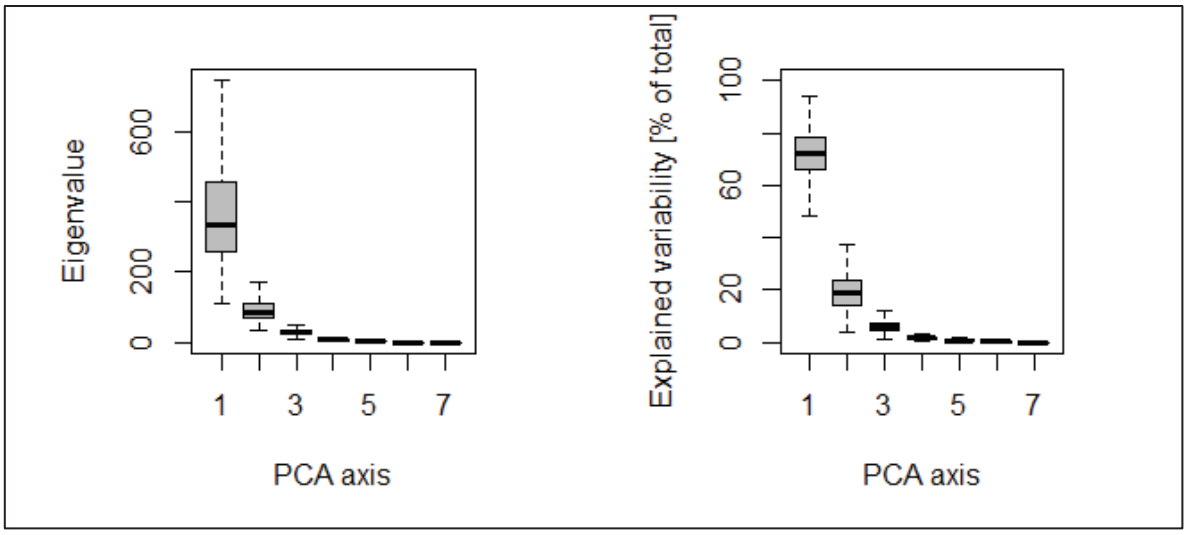
3.3.2.1 Determinación de factores que aportan a la propagación de incendios

En cuanto al índice de propagación de incendios, se plantea determinar los factores que contribuyen a que un incendio se propague; así, la herramienta estadística que se utiliza es el análisis de componentes principales bayesiano (ACPB).

Para este caso, se utilizan las variables correspondientes a la cobertura vegetal, las variables meteorológicas y las variables topográficas correspondientes a los incendios registrados dentro del Parque Metropolitano Guangüiltahua; se deja fuera de este análisis a la cantidad de hectáreas quemadas por los incendios y el tiempo que duró cada uno.

Se realizó un análisis descriptivo de las variables consideradas para poder determinar las que tienen una mayor variabilidad incluida en ellas. De este análisis se determinó que los valores de varianza de la exposición solana y la dirección del viento tenían valores extremadamente grandes en comparación de las demás variables. Para poder disminuir este efecto, se decidió dividir las variables por 100 de tal manera que los valores de la varianza se reduzcan y de esa manera poder realizar un mejor análisis.

Figura 3. Valores propios obtenidos a partir del ACPB



Elaboración: Autores

Tabla 8. Valores propios obtenidos a partir del ACPB

	F1	F2	F3	F4
Mínimo	111,4	33,56	10,07	3,27
Primer cuartil	259,8	68,71	21,72	6,33
Mediana	334,4	86,45	26,91	7,73
Media	379,7	93,3	29,06	8,22
Tercer cuartil	454	110,22	33,88	9,60
Máximo	3110,2	311,64	91,82	23,58

Elaboración: Autores

Tabla 9. Varianza explicada por los valores propios

	F1	F2	F3	F4
Mínimo	45,04	4,04	0,86	0,32
Primer cuartil	66,08	14,16	42,81	12,14
Mediana	71,86	18,67	56,93	16,11
Media	71,72	19,26	61,24	17,55
Tercer cuartil	78,00	23,59	74,55	21,40
Máximo	94,04	43,69	187,27	68,77

Elaboración: Autores

De la figura y la tabla anterior se puede concluir que con dos ejes factoriales se puede explicar más del 90% de la varianza total de las variables; por tanto, se realiza el análisis de los factores mencionados, a continuación se presentan los resultados de las cargas factoriales:

Tabla 10. Cargas Factoriales para el ACPB

	Factor 1	Factor 2
Velocidad del viento	-0,09298507	-0,22684086
Dirección del viento	-0,03653741	-0,06454813
Radiación Solar	-0,10294374	-0,20199756
Temperatura	-0,10409717	-0,23905268
Días sin lluvia	-0,84229706	-0,7475947
Biomasa	-0,75499805	-0,89706582
Pendiente	-0,40264548	-0,76651118

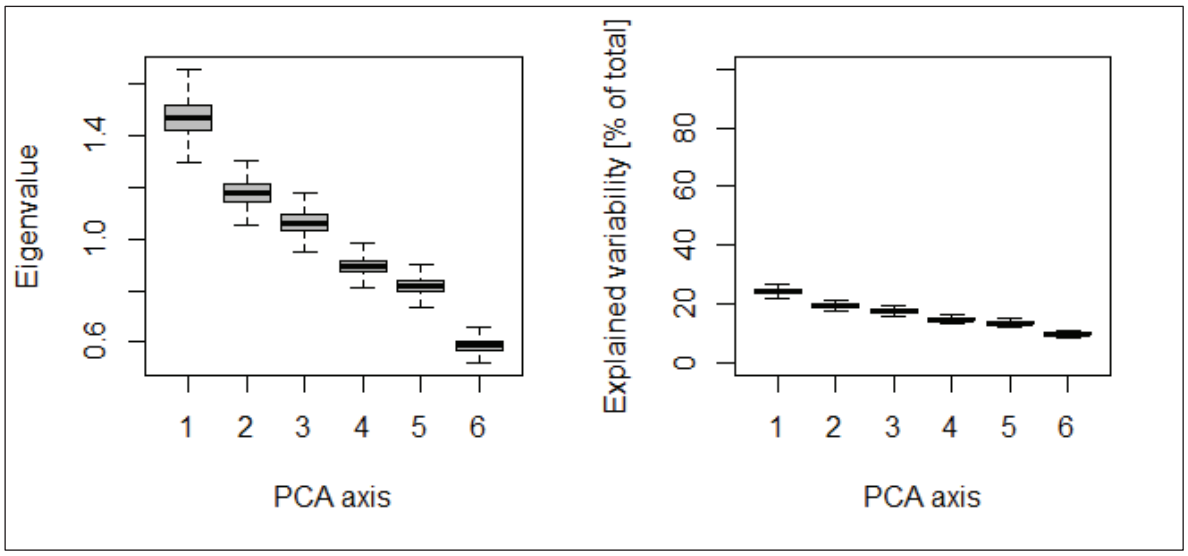
Elaboración: Autores

A través de las cargas factoriales se puede determinar que en el segundo eje factorial se encuentran bien representadas las variables: días sin lluvia, biomasa y pendiente. Por tanto, este eje se denominará: ***Factor de condiciones para la propagación.***

3.3.2.1 Índice de propagación considerando datos de todo el DMQ

Sin embargo, a pesar de que se logró determinar puntuaciones para las variables dentro de los factores, los resultados lastimosamente no son fiables debido a que la cantidad de incendios es baja (13 incendios). Mas, para mostrar que el procedimiento realizado dentro de este proyecto es válido, se va a proceder a determinar los factores del ACPB con la información de todos los incendios registrados en el DMQ; así, se obtiene lo siguiente:

Figura 4. Valores propios obtenidos a partir del ACPB General



Elaboración: Autores

Tabla 11. Valores propios obtenidos a partir del ACPB General

	F1	F2	F3	F4
Mínimo	1,275	1,038	0,9416	0,7847
Primer cuartil	1,424	1,145	1,0341	0,8697
Mediana	1,467	1,176	1,0618	0,8913
Media	1,472	1,178	1,063	0,8938
Tercer cuartil	1,518	1,209	1,092	0,9162
Máximo	1,741	1,346	1,2371	1,0332

Elaboración: Autores

Se realiza el análisis de los valores propios para poder determinar la cantidad de factores a retenerse; no son muy claros los gráficos ni la tabla anteriores; sin embargo, se decide considerar dos factores, que en su valor mínimo son mayores que 1.

Tabla 12. Cargas Factoriales para el ACPB General, 3 Cuartil

	Factor 1	Factor 2
Dirección del viento	0,2635095	0,7672459
Temperatura	0,6064029	0,4107294
Días sin lluvia	0,5884161	0,3443055
Radiación Solar	0,7206376	0,2333826
Biomasa	0,3051184	0,7344169
Pendiente	0,2191064	0,7884273

Elaboración: Autores

Cabe recalcar que se dejó fuera de este análisis a la velocidad del viento debido a que en los dos ejes retenidos las coordenadas de esta variable eran 0 y no aportaban en el modelo. Además, en la tabla anterior se presentan los resultados correspondientes al tercer cuartil del conjunto de análisis (característica del ACPB).

Como se puede ver en la tabla 3.11, se representan bien 3 variables en cada uno de los factores. Por tanto, se tiene: el factor climatológico (Factor 1) y el factor de combustible y condiciones topográficas (factor 2).

3.3.2.2 Predicción del área que ocuparía un incendio de llegar a iniciarse

En esta sección, se busca determinar la cantidad de hectáreas que se quemarían en el caso de que se produjese un incendio. Para este fin, se utiliza una regresión lineal múltiple bayesiana (RLMB), que permite obtener una predicción de las hectáreas quemadas.

Para este modelo se utilizan los factores encontrados en la sección anterior y la velocidad del viento. Adicionalmente, se consideran tres variables *dummy* que son:

$$D_1 = \begin{cases} 1 & \text{si la hectarea quemada es menor que 0,02} \\ 0 & \text{caso contrario} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{si la hectarea quemada es mayor o igual que 0,02 y menor o igual que 2} \\ 0 & \text{caso contrario} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{si la hectarea quemada es mayor que 2 y menor que 20} \\ 0 & \text{caso contrario} \end{cases}$$

El umbral de la primera variable dummy (D_1) fue tomado luego de hacer un análisis descriptivo de los incendios e identificar que existía una concentración grande de puntos por debajo de este valor; de la misma manera se procedió con los umbrales de la segunda y tercera variable *dummy* (D_2 y D_3). Finalmente, el modelo obtenido es:

Tabla 13. Modelo de regresión lineal bayesiana estimado

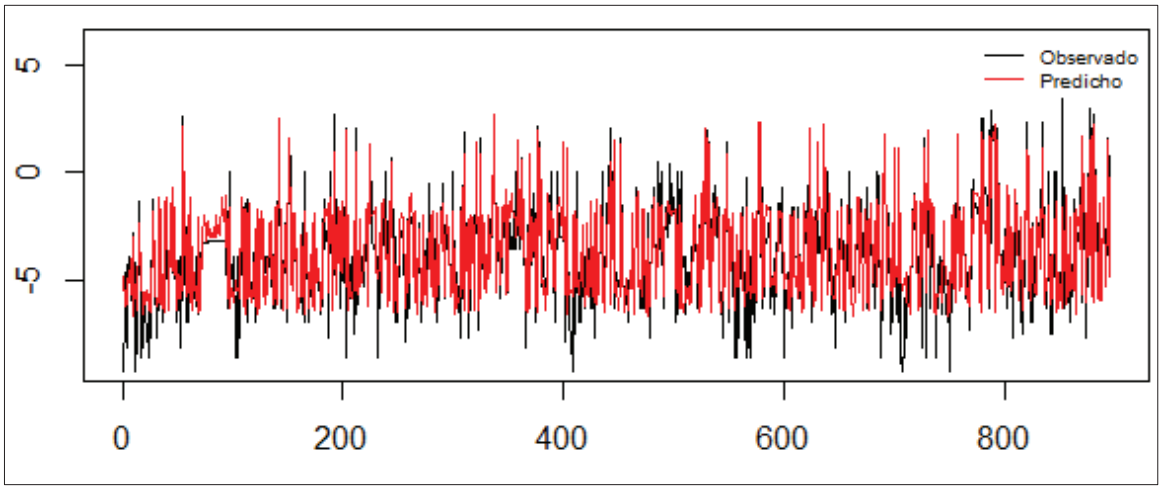
```
Coefficients:
                    Estimate   Std. Error   t value    Pr(>|t|)
log(Vel.viento)    -0.47832    0.05783       -8.271     4.79e-16 ***
log(Factor2)       -0.22414    0.01349      -16.612    < 2e-16 ***
D3                  3.55317    0.19022       18.679    < 2e-16 ***
D1                 -3.49991    0.08842      -39.582    < 2e-16 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 1.764524)
    Null deviance: 17761  on 896  degrees of freedom
Residual deviance:  1574  on 892  degrees of freedom      AIC: 3057.6
```

Elaboración: Autores

Los conceptos referentes a la estimación, los conjuntos creíbles y las pruebas de hipótesis se pueden calcular de manera análoga a la presentada en la sección 3.3.1.1.

Los residuos del modelo tienen una distribución normal. Adicionalmente, se presenta la comparación entre los datos reales y los obtenidos a través del modelo:

Figura 5. Comparación entre los valores reales y estimados en logaritmos para las hectáreas quemadas



Elaboración: Autores

El modelo predice en condiciones extremas lo siguiente:

1. Considerando: Velocidad del viento igual a 0,588; el valor para el factor 2 de 147,936, D1 igual a 0 y la variable D3 igual a 1, se tiene una predicción de hectáreas quemadas igual a 14,687 como valor máximo de la predicción.
2. Considerando: Velocidad del viento igual a 7,7; el valor para el factor 2 de 17.069,14, D1 igual a 1 y la variable D3 igual a 0, se tiene una predicción de hectáreas quemadas igual a 0,00128 como valor mínimo de la predicción.

Se puede observar que son valores que se encuentran dentro de los parámetros encontrados en los incendios del DMQ.

CAPÍTULO 4
ANÁLISIS Y SISTEMATIZACIÓN DE RESULTADOS

De los modelos expuestos en el capítulo anterior, se logró construir y estimar los modelos estadísticos que ayudan a prevenir y analizar la propagación de un incendio.

4.1 ÍNDICE DE PREVENCIÓN DE INCENDIOS

El índice de prevención de incendios se lo estima a partir de la ecuación:

$$\pi_i = \frac{\exp(-5{,}69 + 0{,}24\,T + 0{,}01\,DS)}{1 + \exp(-5{,}69 + 0{,}24\,T + 0{,}01\,DS)}$$

donde,

π_i: es la probabilidad de que un incendio ocurra.
T: es la temperatura medida en grados centígrados.
DS: son los días consecutivos sin lluvia.

Con esta ecuación se logró determinar una probabilidad de incendio en condiciones extremas bajas del 2,4% y una probabilidad de incendio en condiciones extremas altas de 84,14%. Así, el índice de prevención de incendios, queda determinado de la siguiente manera:

Tabla 14. Índice de prevención de incendios

Riesgo	Probabilidad
Bajo	Menor o igual que 20,44%
Medio	De 20,45% a 40,87%
Alto	De 40,88% a 61,31%
Grave	Mayor que 61,31%

Elaboración: Autores

4.2 ÍNDICE DE PROPAGACIÓN DE INCENDIOS

En el caso del índice de propagación de incendios se realizó dos análisis, el primero para determinar posibles factores dentro de las variables que se están analizando y por otro una predicción de las hectáreas quemadas.

Se encontraron dos factores, cuyas ecuaciones se describen a continuación:

$$F1 = 0{,}61\ T + 0{,}59\ DS + 0{,}72\ RS$$

$$F2 = 0{,}77\ DV + 0{,}73\ B + 0{,}79\ P$$

donde,

$F1$: es el factor 1.
$F2$: es el factor 2.
T: es la temperatura medida en grados centígrados.
DS: son los números consecutivos sin lluvia.
RS: es la radiación solar.
DV: es la dirección del viento.
B: es la biomasa.
P: es la pendiente del terreno.

Luego, se precedió a estimar un modelo de regresión para poder determinar la cantidad estimada de hectáreas que se quemarían si un incendio llegara a producirse. El modelo resultante es el siguiente:

$$\log(H) = -0{,}48 \log(VV) - 0{,}22 \log(F2) + 3{,}55\ D3 - 3{,}50\ D1$$

donde,

H: son las hectáreas quemadas.
$F2$: es el factor 2.
$D1$: es la variable *dummy,* que tiene valor 1 si el área quemada es menor que 0,002 hectáreas.
VV: es la velocidad del viento.
$D3$: es la variable *dummy,* que tiene valor 1 si el área quemada es mayor que 2 y menor que 20 hectáreas.

Con esta ecuación se logró determinar las hectáreas que se verían afectadas en caso de ocurrir un incendio, en condiciones extremas bajas de 0,00128 hectáreas y en condiciones extremas altas un total de 14,687 hectáreas quemadas. De esta manera, el índice de propagación es:

Tabla 15. Índice de prevención de incendios

Riesgo de propagación	Hectáreas quemadas
Bajo	Menor o igual que 0,73
Medio	De 0,74 a 1,47
Alto	De 1,48 a 3,67
Grave	Mayor que 3,67

Elaboración: Autores

4.3 SISTEMATIZACIÓN DE LOS RESULTADOS

En esta sección, se presentan algunas simulaciones de los índices obtenidos por los modelos precedentes en una interfaz gráfica; de tal manera que se pueda visualizar claramente los resultados obtenidos en este proyecto.

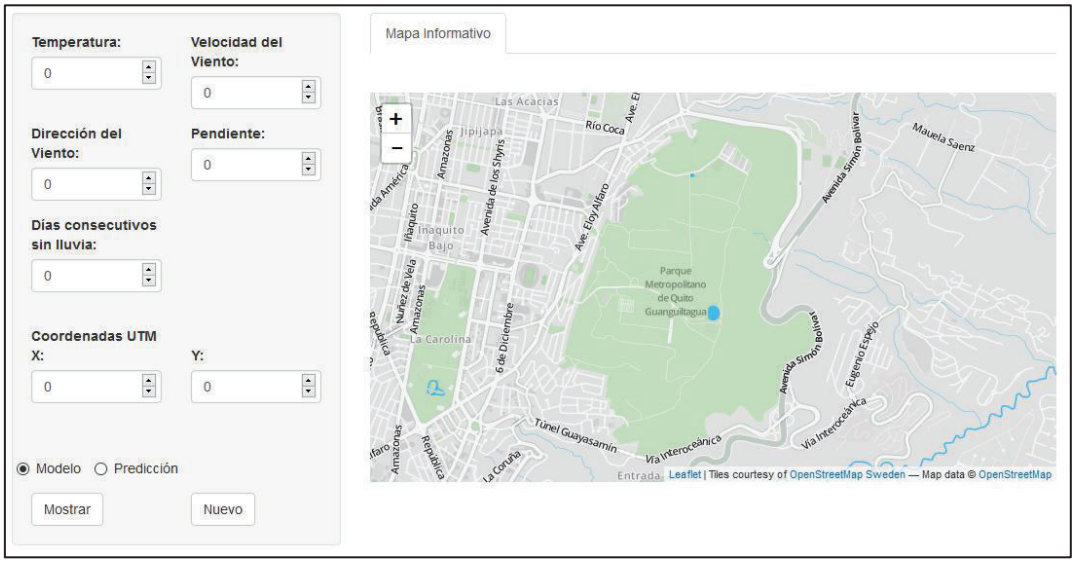
En este caso, se ha utilizado la aplicación *Shiny*, del *software* R. Esta herramienta permite la construcción de aplicativos web y en particular de mapas con la ayuda de *R Project,* para poder visualizar los resultados en un aplicativo.

Esta aplicación fue diseñada para que muestre; por un lado un mapa interactivo de los modelos de los índices para los 14 incendios del parque metropolitano, y por otro lado permita introducir parámetros para obtener predicciones de los índices y visualizarlos automáticamente. Además, la aplicación se estructuró de tal manera que se puede introducir los parámetros para una predicción, seleccionar entre el mapa del modelo o el de la predicción y ejecutar el mapa respectivo. Por otro lado, se visualiza el mapa con la información del modelo o de la predicción.

Cuando se inicia la aplicación cada parámetro se inicializa en 0, está seleccionada por defecto la opción modelo y se visualiza un mapa del parque metropolitano. Cabe recalcar que los mapas se visualizan de manera dinámica, es decir, podemos navegar en ellos y acercarnos o

alejarnos de los mismos para tener una mejor percepción de lo obtenido, lo cual no se puede hacer con mapas estáticos. A continuación se observa la aplicación una vez iniciada:

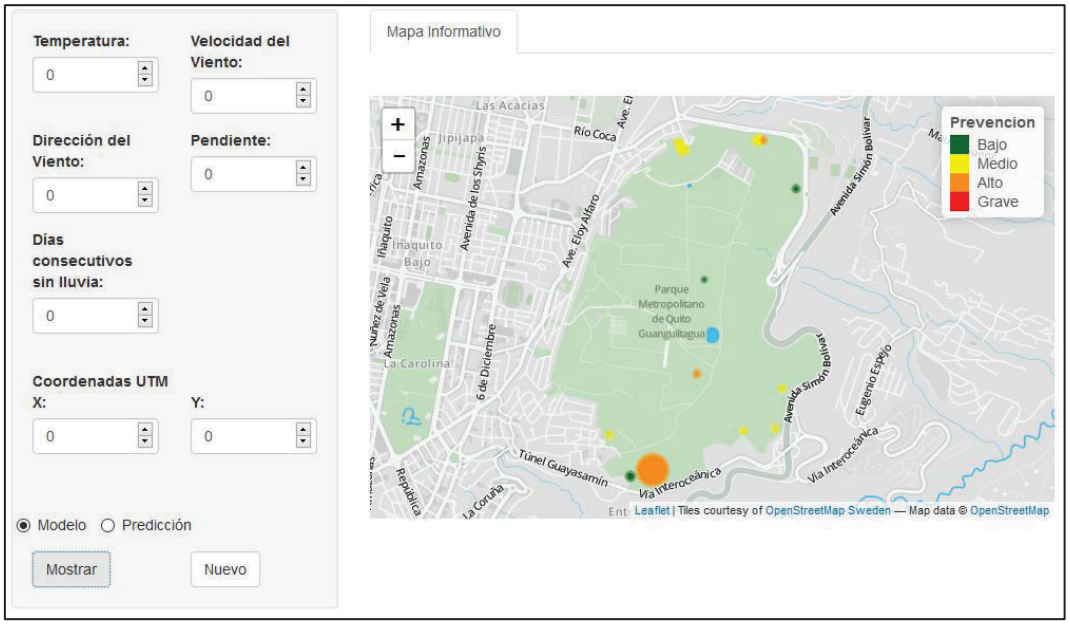
Figura 6. Inicio de la aplicación



Elaboración: Autores

Para visualizar los resultados de los índices del sistema elaborado, se selecciona la opción modelo y se da clic sobre el botón mostrar; así, se visualizan los 14 incendios ocurridos en el parque dentro del mapa con el índice de prevención asociado a cada incendio mediante su respectivo color (bajo=verde, medio=amarillo, alto=naranja y grave=rojo) como se muestra a continuación:

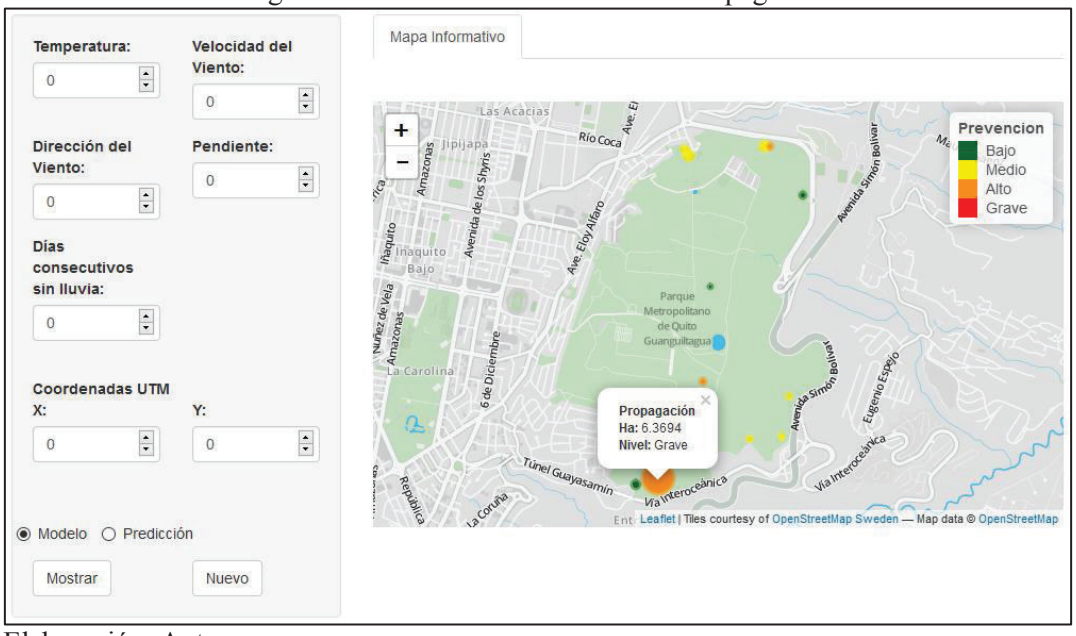
Figura 7. Visualización del Índice de Provención



Elaboración: Autores

Para obtener la información sobre el índice de propagación basta dar un clic sobre cada uno de los incendios y se desplegará una caja de texto, la misma que índica las hectáreas consumidas por el incendio y el nivel asociado al índice (bajo, medio, alto, grave), como se muestra en la siguiente figura:

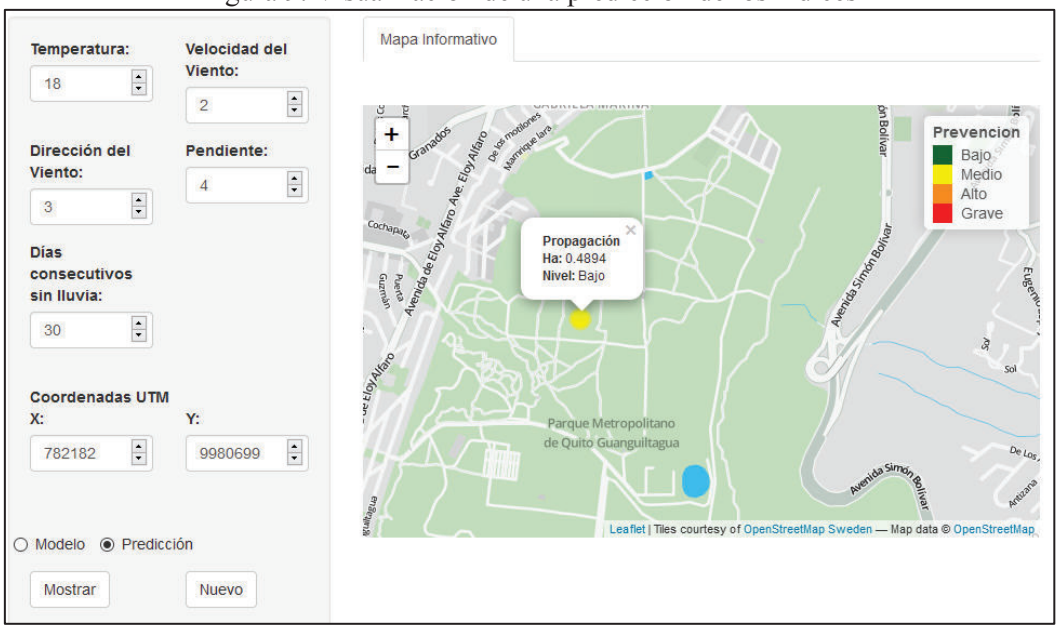
Figura 8. Visualización del Índice de Propagación



Elaboración: Autores

Para mostrar una predicción sobre los índices, se deben llenar cada uno de los parámetros del panel izquierdo, luego se procede seleccionando la opción ***Predicción*** y; finalmente, para visualizar el resultado se debe dar clic sobre el botón ***mostrar***, en el panel derecho aparecerá el incendio con su respectivo índice de prevención y de manera similar al dar clic sobre este aparecerá la información correspondiente al índice de propagación.

Figura 9. Visualización de una predicción de los índices



Elaboración: Autores

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

- Lastimosamente fue imposible realizar la estimación del SIPIF con los datos del parque metropolitano Guangüiltagua ya que la cantidad de datos no era suficiente para estimar los modelos propuestos en este proyecto por el mal estado de las bases de datos. Sin embargo, al considerar los incendios forestales del DMQ y adaptarlos a las condiciones climáticas del parque metropolitano, fue posible construir y estimar los índices propuestos en este proyecto.

  Es decir, si se tuvieran una cantidad grande datos bien validados y coherentes, es posible que la metodología utilizada en este documento y los resultados obtenidos tengan una mayor validez estadística. Sin embargo, este proceso tomaría un trabajo importante de varios profesionales en temas de biología, meteorología, estadística, entre otros.

- Los índices estimados en este proyecto tienen consistencia con la realidad y con los datos registrados de incendios en el DMQ. Se constituye entonces como una herramienta de apoyo en la prevención de los incendios y mitigar los daños causados por estos. Los resultados de este proyecto pueden utilizarse como una línea base para modelos de estimación de índices que cuenten con mejores datos y sean más elaborados.

- Se agregaron variables que antes no se han considerado en el país para la elaboración de los índices, en especial en el caso del índice de propagación, como son: la dirección del viento, la pendiente del terreno y la cantidad de vegetación medida en hectáreas de cobertura. Además, se puede agregar más valor al modelo el momento que se realice una construcción sistemática desde la base de datos.

- Con los resultados encontrados y la ayuda del aplicativo *Shiny* del paquete estadístico *R Project*, se diseñó un mapa georeferenciado de prevención y propagación de los incendios en el parque metropolitano. Con esto se puede visualizar los posibles focos de incendio y la cantidad de terreno que cubrirían.

RECOMENDACIONES

- Se recomienda mejorar la toma de información de los eventos relacionados con incendios, debido a que actualmente las bases de datos no poseen información suficientemente fidedigna.

- Ya que este proyecto sirve de línea base para la prevención de incendios, se recomienda realizar un estudio más elaborado y profundo sobre la temática de tal manera que se pueda aplicar a todo el DMQ.

- En un futuro, para robustecer este proyecto se debe contar con un equipo completo de profesionales en diferentes áreas, por ejemplo: geógrafos, meteorólogos, ingenieros ambientales, estadísticos, expertos en incendios, etc.

- Se debe tener un mejor control sobre los equipos y personas que miden las variables meteorológicas, ya que dentro del análisis de este proyecto, se encontraron datos perdidos, llegando incluso a no tener todo un año de medición.

- Se recomienda además, investigar sobre la combustibilidad de cada uno de los tipos de coberturas vegetales del DMQ, de tal manera que esta variable ayude a la mejor estimación de futuros modelos.

BIBLIOGRAFÍA

INCENDIOS

Aguado, I., Camia, A., **"Fundamentos y utilización de índices meteorológicos de peligro de incendio"**, Serie Geográfica Vol. 7, pág. 49 – 58, España, 1998.

Álvarez, M., De Santis, A., Chuvieco, E., **"Estimación del peligro de incendios a partir de teledetección y variables meteorológicas: variación temporal del contenido de humedad del combustible"**, Universidad de Santiago de Compostela, España, 2005.

Badia, A, y otros, **"prevención de incendios forestales. Integración de las técnicas de modelización en los sistema de información geográfica"**, Universitat Autónoma de Barcelona, España, 1997.

Berry, DA, **"Statistics: A Bayesian perspective"**. Duxbury Press, Belmont, California, 1996.

Dentoni, M. y Muñoz M., **"Evaluación de peligro de incendios. Informes técnicos. Informe Técnico Nº1. Sistemas de Evaluación de Peligro de Incendios",** Chubut, Argentina, 2012.

Dirección General de Protección Civil y Emergencias, **"VADEMECUM REMER",** España, 2014.

Julio, G., **"Diseño de índices de riesgo de incendios forestales para Chile"**, Instituto de manejo forestal, Chile, 1990.

Natural Resourse Canada, **"Canadian Forest Fire Danger Rating System (CFFDRS)"**, Canadá, 2008.

Sala de Situación Metropolitana, EMSEGURIDAD, **"Informe de Situación: PLAN DE PREVENCIÓN Y RESPUESTA PARA INCENDIOS FORESTALES DEL DMQ. Período: 26 de junio al 7 de septiembre de 2014",** Ecuador, 2014.

Setzer, A., Sismanoglu, R., **"Risco de Fogo: Metodologia do Cálculo – Descricao sucinta da Versao 9"**, Instituto Nacional de Pesquisas Espaciais, Brasil, 2012.

Sistemas evaluación de riesgo de incendio forestal. Miliarium.com, Monografías, Incendios forestales, 2008.

Torres, J., Magaña, O., Ramírez, A., **"Índice de peligro de incendios forestales a largo plazo"**, Centro de Investigación y Docencia Económica, México, 2006.

ESTADÍSTICA BAYESIANA

Albert, J., Chib, S., **"Bayesian Análisis of Binary and Polychotomous Response Data"**, Journal of the American Statistical Association, Vol. 88. No. 422, pág. 669 – 679, 1993.

Albert, J., Chib, S., **"Bayesian residual analysis for binary response regression models"**, Biometrika, Vol. 82, No. 4, pág. 747 – 759, Great Britain, 1995.

Bazán, J., Bayes, C., **"Inferencia Bayesiana en modelos de regression binaria usando BRMUW"**, Pontifica Universidad Católica del Perú, Perú, 2010.

Congdon, P., **"Bayesian Statistical Modelling"**, second edition, Willey Series in Probability and Statistics, England, 2006.

Knight, K., **"Mathematical Statistics",** Chapman & Hall, United Estates, 2000.

Nounou, M., Bakshi, B., Goel, P., Shen, X., **"Bayesian principal component analysis"**, Journal of Chemometrics, Vol. 16, pág. 576 – 595, 2002.

O'Brien, S., Dunson, D., **"Bayesian Multivariate Logistic Regression"**, Biometrics, Vol. 60, pág. 739 – 746, 2004.

Rezghi, M., Obulkasim, A., **"Noise – Free principal component analysis: An efficient dimension reduction technique for high dimensional molecular data",** Expert Systems with applications, Vol 41, publicación 17, pág., 7797-7804, 2014.

Sivia, D., Skilling, J., **"Data Analysis. A Bayesian Tutorial"**, Oxford University Press, United States, 2006.

Van Erp, N., Van Gelder, P., **"Bayesian Logistic Regression Analysis"**, AIP Conference Proceedings 1553, 147, 2013.

PAQUETE R

R Core Teem, **"R: An Language and Environment for Statistical Computing"**, R Foundation for Statistical Computing, Austria, 2015.

ANEXO. CONCEPTOS ADICIONALES

1. Distribuciones

a) Distribución χ^2 no centrada

Si $X_1, X_2, X_3, \dots, X_n$ son variables independientes y $X_i \sim N(\mu_i, 1)$ para $i = 1, \dots, n$, entonces

$$\sum_{i=1}^{n} X_i^2 \sim \chi_{n,\lambda}^{2'}$$

Donde $\chi_{n,\lambda}^{2'}$ es la distribución χ^2 no centrada, con n grados de libertad y parámetro $\lambda = \sum_{i=1}^{n} \mu_i^2$ de no centralidad.

b) Distribución Gamma-Inversa

La variable aleatoria continua X tiene una distribución gamma-inversa $(GInv)$ con parámetros $a > 0, b > 0$ si:

$$GInv(x|a,b) = cx^{-(a+1)}e^{-b/x}, \quad x > 0$$

donde,

$$c = \frac{b^a}{\Gamma(a)}$$

$$E[x] = \frac{b}{a-1}, \qquad a > 1$$

$$V[x] = \frac{b^2}{(a-1)^2(a-2)}, \qquad a > 2$$

c) Distribución t-Student multivariada

Un vector aleatorio $X = (X_1, \dots \dots, X_k)'$ tiene una distribución t-Student multivariada de dimensión k, con parámetros $\mu = (\mu_1, \dots \dots, \mu_k)'$, M y a ($\mu \in \mathbb{R}^k$, M es una matriz de tamaño $k \times k$ simétrica definida positiva, $a > 0$), si

$$t_k(x|\,\mu, M, a) = c\,\left[1 + \frac{1}{a}(x-\mu)^t M\,(x-\mu)\right]^{-(a+k)/2}, \qquad x \in \mathbb{R}^k$$

donde,

$$c = \frac{\Gamma[(a+k)/2]}{\Gamma[a/2](a\,\pi)^{k/2}}|M|^{1/2}$$

$|M|$: es el determinante de la matriz M.

2. Integrales relacionadas con la distribución gamma-inversa

Sea η un vector de constantes de $n \times 1$ y C una matriz definida positiva y simétrica, $x \in \mathbb{R}$.

$$\int_{-\infty}^{\infty} \exp\left[\frac{1}{2}(x-\eta)^t C^{-1}(x-\eta)\right] dx = \sqrt{2\pi}\,|C|^{1/2}$$

Para $a > 0\,, p > 0\,, x > 0$

$$\int_0^{\infty} x^{p-1} e^{-ax^2} dx = \frac{1}{2} a^{-\frac{p}{2}}\,\Gamma(p/2)$$

3. Teoremas

I. **Teorema 1.** Sean A una matriz de $r \times r$ de constantes y Y un vector aleatorio de $r \times 1$, con media μ y una matriz no singular de V de covarianza, entonces:

- $Var[AY] = AVA'$
- $E[Y'AY] = tr(AV) + \mu' A\mu$
  Si $V = \sigma^2 I$, entonces $E[Y'AY] = \sigma\, tr(A) + \mu' A\mu$

II. **Teorema 2.** Sean A una matriz de $r \times r$ de constantes y Y un vector aleatorio normal multivariado de $r \times 1$, con media μ y matriz no singular $\sigma^2 I$ de covarianza ($Y \sim N(\mu, \sigma^2 I)$).

Sea U la forma cuadrática dada por $U = Y'AY$.

Si A es idempotente y de rango q; entonces

$$\frac{U}{\sigma^2} \sim \chi^{2'}_{q,\lambda}$$

donde,

$$\lambda = \mu' A \mu / \sigma^2$$