Análisis geoestadístico de datos de la contaminación del aire en Santiago de Chile, usando SPDE con método de estimación INLA.

Tesis presentada para optar al grado de Magíster en Estadística.

Estudiante: Eduardo Puraivan H.
Profesora guía: Orietta Nicolis.

Becario CONICYT

Agradecimientos

Gracias a todas las personas que son su cariño, comprensión y apoyo han permitido que cada día pueda aprender algo más de la vida y la ciencia. Con mucho cariño dedico a cada uno de ustedes:

Karina, Javiera y Valeria gracias por acompañarme en mis sueños.

Profesora Orietta, gracias por sus enseñanzas y sus sabios consejos.

A los profesores del DEUV agradezco sus enseñanzas y buena voluntad.

Enner, gracias por tu paciencia y por enseñarme tu humildad.

Además agradecer a CONICYT, sin el apoyo económico de las becas muchos no podríamos lograr nuestros sueños.

Becario CONICYT.

A mis hijas; Javiera y Valeria.

Índice general

Índice general III

Índice de figuras VII

Índice de cuadros VIII

1 Introducción 2

2 Contexto del estudio 4

2.1. Problema de investigación 4

2.1.1. Objetivos de investigación 4

2.2. Material particulado 5

2.2.1. Material particulado atmosférico 5

2.2.2. PM_{10} y $PM_{2,5}$ 6

2.2.3. Normas internacionales 9

2.2.4. Normativa legal en Chile 9

2.2.5. Estación de monitoreo en Santiago de Chile 10

3 Conceptos de Geoestadística 11

3.1. Introducción 11

3.2. Definición de geoestadística y campo de estudio 12

3.3. Conceptos básicos de geoestadística 12

3.3.1. Variable regionalizada 12

3.3.2. Estacionariedad 13
3.3.3. Funciones de correlación espacial 14
3.3.3.1. Variograma y semivariograma 15
3.3.3.2. Covarianza y correlograma 16
3.3.3.3. Isotropía 17
3.3.4. Covarianza y variogramas 17
3.3.5. Función de covarianza 18
3.4. Predicción espacial clásica 20
3.5. Estadística Bayesiana 23
3.5.1. Conceptos básicos 23
3.5.2. Teorema de Bayes 23
3.5.3. Inferencia Bayesiana 24
3.5.4. Distribución a priori 24
3.5.5. Función de verosimilitud 24
3.5.6. Distribución a posteriori 25
3.6. Campos Gaussianos y Campos Aleatorios de Markov Gaussianos 25
3.7. Enfoque SPDE 26

4 Modelo espacio temporal 27

4.1. Modelo espacio temporal. 27
4.2. GMRF y SPDE 30
4.3. Enfoque SPDE 31
4.4. El modelo en R. 32

5 Aplicación del modelo a datos de la contaminación del aire en Santiago de Chile. 34

5.1. Región de estudio y variables 34
5.1.1. Región de estudio desde GoogleMap 34
5.1.2. Contorno de la región de estudio con R 35
5.1.3. Estaciones de monitoreo. 36

5.1.4. Base de datos del $PM_{2,5}$ . . . 36
5.2. Análisis exploratorio. . . . 37
5.2.1. Resumen del $PM_{2,5}$ en las estaciones de monitoreo. . . . 37
5.2.2. Box plot del $PM_{2,5}$ . . . 38
5.2.3. Serie de tiempo de los datos y autocorrelación. . . . 38
5.2.4. Comportamiento temporal del $PM_{2,5}$ . . . 39
5.2.5. $PM_{2,5}$ en cada mes . . . 41
5.2.6. Resumen anual . . . 41
5.2.7. Covariables . . . 42
5.2.8. Comportamiento temporal de la Temperatura. . . . 42
5.2.9. Comportamiento temporal de la Humedad Relativa . . . 43
5.2.10. Comportamiento temporal de la velocidad del viento . . . 44
5.2.11. Modelo lineal simple . . . 44
5.3. Resultados finales . . . 45
5.3.1. Base de datos (DATOS) . . . 45
5.3.2. Base de datos de validación (DATOS_VAL) . . . 46
5.3.3. Densidad del PM. . . . 47
5.3.4. Triangulación de la región: . . . 48
5.3.5. Definición del Modelo . . . 49
5.3.6. Parámetros del Modelo . . . 51
5.3.7. Mapa de estimación del $PM_{2,5}$ . . . 52
5.3.8. Conclusiones . . . 53

6 Apéndice. 55

6.1. Código de Tesis . . . 55
6.1.1. Análisis construcción de base de datos . . . 55
6.1.2. Modelo en R . . . 57
6.2. Método de Laplace . . . 62
6.2.1. Ejemplo binomial . . . 63

6.3. Ejemplos de R -inla 65
6.3.1. Modelo binominal 65
6.3.2. SPDE. Triangulación 66
6.3.2.1. Ejemplo SPDE 67

Bibliografía **70**

Índice de figuras

2.2.1.Estación de monitoreo, Santiago de Chile 10

5.1.1.Santiago de Chile, vista satelital. 35
5.1.2.Contorno de Santiago de Chile con estaciones de monitoreo. 35
5.1.3.Estructura de la base de datos. 37
5.2.1.box-plot del $PM_{2,5}$ por estación según año. 38
5.2.2.Autocorrelación temporal 39
5.2.3.Nivel del $PM_{2,5}$ por día según estación. 40
5.2.4.Nivel del $PM_{2,5}$ por mes según año. 41
5.2.5.Nivel del $PM_{2,5}$ anual 42
5.2.6.Temperatura por estación según año. 43
5.2.7.Humedad relativa por estación según año. 43
5.2.8.Velocidad del viento por estación según año. 44
5.3.1.Estructura de la base de datos. 46
5.3.2.Estructura de la base de datos de validación. 46
5.3.3.Estructura de la base de datos (estandarizada) 47
5.3.4.Densidad $PM_{2,5}$ y $log(PM_{2,5})$ 47
5.3.5.Triangulación con INLA. 49
5.3.6.Media y desviación estándar 53

6.3.1.Gráficas de triangulación. 68

Índice de cuadros

2.1. Caracterización de las fracciones del material particulado presente en el aire troposférico. 6
2.2. Impactos generados por $MP, O_3, SO_2 \quad y \quad NO_2$ 8
2.3. Niveles del material particulado (OMS) 9
2.5. Niveles del $PM_{2,5}$ (leyChile) 10

5.1. Estaciones de monitoreo (Santiago de Chile) 36
5.3. Resumen del $PM_{2,5}$ por estación 37
5.4. Número de días que se supera la norma por estación, según año. 40
5.6. Coeficientes del modelo 45
5.7. Parámetros del modelo 51
5.8. Hiperparámetros 52
5.9. Ejemplo de Mapas. 52

Capítulo 1

Introducción

Hablar de estadística resulta familiar, al menos sabemos que en su esencia permite estudiar datos para posteriormente tomar decisiones, ya sea en el área de la ingeniería, educación, familia, presupuesto, medio ambiente entre otras áreas.

En Chile el currículo educativo nos presenta dentro de la asignatura Educación Matemática, un eje llamado Datos y Azar. Es allí donde reconocemos la noción de probabilidad, desde la óptica frecuentista en una primera instancia, de esta forma vamos adoptando un vocabulario que ya nos permite hablar de experimentos aleatorios, muestra, eventos, axiomas de Kolmogorov, independencia de eventos, teorema de Bayes, entre otros temas.

Si bien los ejemplos dados por nuestros profesores(as) nos permiten avanzar, parecen alejados de la realidad y más de alguna vez nos preguntamos: ¿y cuándo aplicamos todo esto?

Lo cierto es que la estadística permite a la comunidad científica, personal académico y en general a todas las personas analizar diferentes fenómenos de interés. Para ello la estadística cuenta con diversos resultados teóricos que van desde hacer un conteo para ver la frecuencia de ocurrencia de un determinado evento hasta estudiar modelos estadísticos y otros temas avanzados.

En esta oportunidad se ha optado por estudiar un fenómenos muy preocupante y cada vez más frecuente en nuestro país, *la contaminación atmosférica*. Santiago, Temuco, Los Angeles son algunas de nuestras ciudades donde frecuentemente existe alerta ambiental, el efecto de la

contaminación implica diversos daños no solamente al ser humano sino que a todo el ecosistema.

El basar este trabajo en la contaminación tiene dos razones fundamentales:

- La contaminación atmosférica provoca daño grave a la salud y por lo tanto el tema sin dudas tiene gran relevancia.
- Los métodos utilizados desde la estadística resultan muy interesantes, estos van desde analizar un gran número de datos, hasta preguntarnos cómo debemos modelar e inferir.

Desde la óptica de la estadística el trabajo tiene como propósito fundamental implementar el método de inferencia numérica INLA (Integregrated Nested Laplace Aproximation) introducido por [Rue et al (2009)] , método que en conjunto con SPDE (Stochastics Partial Diferential Equation) propuesto por [Lindgren et al (2011)] se utilizarán para estudiar datos ambientales en la ciudad de Santiago de Chile, específicamente el $PM_{2,5}$.

La organización del trabajo es mediante capítulos:

- Capítulo I: Introducción
- Capitulo II: Contexto del estudio
- Capítulo III: Conceptos de Geoestadística
- Capítulo IV: Modelo espacio temporal
- Capítulo V: Aplicación del modelo a datos de la contaminación del aire de la ciudad de Santiago de Chile.

Capítulo 2

Contexto del estudio

2.1. Problema de investigación

En este trabajo se estudia un modelo espacio-temporal para modelar el $PM_{2,5}$ en función de la temperatura, humedad relativa y velocidad del viento, además de componentes espaciales. El modelo se estudia bajo un enfoque de inferencia bayesiana usando método INLA.

En concreto se explora un modelo propuesto por [Rue et al (2009)] para el estudio de la contaminación atmosférica en la ciudad de Santiago de Chile. Los datos fueron medidos en 11 estaciones de monitoreo durante los años 2012 - 2013 y 2014.

2.1.1. Objetivos de investigación

La investigación persigue los siguientes objetivos:

- Explorar la correlación espacial y temporal de los datos.
- Ajustar un modelo espacio temporal para modelar el $PM_{2,5}$.
- Estimar los parámetros e hiperparámetros del modelo bajo enfoque bayesiano usando método INLA.
- Construir mapa de estimación.

2.2. Material particulado

El material particulado atmosférico se define como un conjunto de partículas sólidas y/o líquidas (a excepción del agua pura) presentes en suspensión en la atmósfera. Existen diversos estudio que señalan los daños hacia la salud y el medio ambiente provocados por la contaminación atmosférica.

2.2.1. Material particulado atmosférico

El material particulado es una mezcla de componentes con características físicas y químicas muy diversas que están determinadas por los mecanismos de sus génesis. Este aerosol se forma a partir de partículas directamente emitidas a la atmósfera (particulado primario) o aquellas que se forman a partir de procesos de conversión gas-partículas (particulado secundario).

Así existe material particulado fino y material particulado grueso. El particulado grueso comprende polvo en suspensión o resuspendido de los caminos y de procesos industriales, construcción, minería, entre otras y también un componente biológico en que destacan el polen y fragmentos de bacterias.

A nivel urbano el desgaste de neumático, frenos y pavimento producen un material particulado del tráfico vehicular. En los sectores rurales la actividad agrícola, la minería y el polvo proveniente de caminos no pavimentados junto a la acción del viento sobre la corteza terrestre adquiere relevancia.

La siguiente tabla extraída de [Vargas (2011)], muestra la caracterización de la fracción del material particulado presente en el aire troposférico.

Cuadro 2.1: Caracterización de las fracciones del material particulado presente en el aire troposférico.

	Material particulado fino($PM_{2,5}$)	Material particulado grueso(PM_{10})
Se forma a partir de	Gases	Sólidos grandes.
Se forma a través de	Reacciones químicas o vaporización. Nucleación, condensación sobre núcleos y coagulación. Evaporación de gotitas de neblina y nubes en que se han disuelto gases.	Disrupción mecánica (aplastamiento, molienda, abrasión de superficies, etc). Evaporación de sprays. Suspensión de polvos.
Están compuestas de	Sulfato, nitrato, amonio, carbono elemental. Compuestos orgánicos como los HAP. Metales como plomo, cadmio, vanadio, niquel, cobre, zinc, manganeso, hierro.	Polvo resuspendido del suelo y las calles. Ceniza del carbón y petróleo. Óxido de elementos de corteza (sílice, aluminio, titanio y hierro). Sal, carbonato de calcio, polen, esporas de hongos, moho. Fragmentos de plantas y animales. Detribus del desgaste de los neumáticos.
Solibicidad	Predominantemente solubles, higroscópico y delicuescente.	Predominantemente insolubles y no higroscópico.
Fuentes	Combustión del carbón, petróleo, gasolina, disel o madera.	Resuspensión del polvo industrial y del suelo en carreteras y calles. Suspensión del suelo en minería, caminos no pavimentados.
Vida media en la atmósfera.	Días a semanas.	Minutos a horas.
Distancia de viaje.	100 a 1000 km.	1 a 10 km.

2.2.2. PM_{10} y $PM_{2,5}$

Actualmente se distinguen al menos dos tipos de material particulado, el material particulado fino y el grueso. A su vez el material particulado fino contiene al material particulado ultra fino con diámetros aerodinámicos menor a 0.1 micrón y que constituyen en número la mayor parte de las partículas .

El diámetro de las partículas en suspensión varía desde nanómetro (nm) hasta la decena de micras (μm). La clasificación de la partícula también puede depender del campo de estudio. Por ejemplo se denomina partícula fina en ciencias atmosféricas a aquellas partículas de diámetro menor a 1 μm mientras que en epidemiología esta definición abarca hasta partículas de diámetro

menor que $2{,}5\mu m$. De la misma manera se denomina partícula gruesa en ciencias atmosféricas a aquellas partículas de diámetro a partir de 1 μm mientras que en epidemiología esta definición abarca hasta partículas de diámetro mayor que $2{,}5\mu m$. .

En términos de calidad del aire se definen cuatro parámetros fundamentales atendiendo al tamaño de corte del sistema de captación: SPT, PM_{10}, $PM_{2{,}5}$ y PM_1. Mientras el término PST se refiere a partículas en suspensión totales, PM_{10} se define como el conjunto de partículas que atraviesa un cabezal de tamaño selectivo para un diámetro aerodinámico de $10\mu m$ con una eficiencia de corte del $50\,\%$. La misma definición para cabezales de corte de $2{,}5\mu m$ y $1\mu m$ se aplica para $PM_{2{,}5}$ y PM_1 respectivamente.

Así debemos considerar que el $PM_{2{,}5}$ se produce por emisiones directas de los procesos de combustión de fósiles, a partir de la condensación de gases, de reacciones químicas en la atmósfera a partir de gases precursores como el dióxido de azufre, óxido de nitrógeno, compuesto orgánicos volátiles, amoníaco y otros compuestos; y a través de proceso de nucleación y coagulación de partículas ultrafinas.

De este modo, las principales fuentes de $PM_{2{,}5}$ son los automóviles, buses, camiones, plantas termoeléctricas, calderas, procesos industriales, hornos, fundiciones, procesos metalúrgicos, calefacción residencial a leña, quemas agrícolas y emisiones de amoníaco de las operaciones agrícolas.

Existen diversos estudios internaciones sobre los efectos adversos hacia la salud. Por ejemplo, enfermedades pulmonares, enfermedades cardiovasculares, aumento de riesgo de infarto entre otras.

La siguiente tabla tomada de [MMA (2011)], muestra los efectos del material particulado.

Cuadro 2.2: Impactos generados por $MP, O_3, SO_2 \quad y \quad NO_2$

Efecto	Descripción
Daño a la salud	Las partículas y compuestos emitidos al aire pueden producir efectos nocivos en la salud de las personas como por ejemplo; reducción de la función pulmonar, aumento de la susceptibilidad de contraer infecciones respiratorias, muerte prematuras y cáncer, entre otras.
Disminución de la visibilidad	La presencia de partículas en el aire reduce la visibilidad causando una disminución en el bienestar y la calidad de vida.
Daños materiales	El exceso de contaminación atmosférica puede causar daños en los materiales de construcción, alterando las propiedades físicas y químicas de las mismas.
Daño a ecosistema acuático	Altas concentraciones de NO_2 y SO_2 pueden producir deposiciones ácidas en el agua, modificando su composición y dificultando la supervivencia de especies acuáticas.
Daño en plantas y Bosques.	La deposición ácida de suelos puede alterar el crecimiento de plantas y árboles. Además el ozono y otras partículas pueden ingresar a través de los estomas de la planta y dañar su estructura.

Cabe señalar que la fracción fina $PM_{2,5}$, está compuesta por partículas pequeñas que penetran la vía respiratoria hasta llegar a los pulmones y los alveolos, lo que aumenta el riesgo de mortalidad prematura por efectos cardiopulmonares [CONAMA (2010)].

2.2.3. Normas internacionales

Referente al material particulado, la Organización mundial de la salud (OMS) señala en su página oficial[1] los siguientes valores de referencia:

Cuadro 2.3: Niveles del material particulado (OMS)

	$PM_{2,5}$	PM_{10}
Índice de media anual	$10\mu g/m^3$	$20\mu g/m^3$
Índice de media 24hrs.	$25\mu g/m^3$	$50\mu g/m^3$

2.2.4. Normativa legal en Chile

De acuerdo con la ley Chilena de bases generales del medio ambiente, es deber del estado diseñar normas para regular la presencia de contaminantes en el medio ambiente, de manera de prevenir que éstos puedan significar o representar, por sus niveles, concentraciones y periodos, un riesgo para la salud de las personas.

La calidad del aire tiene repercusiones directa sobre la población. Si bien en Chile han existido instrumentos y políticas que intentan regular la calidad del aire, el país no cumple con estándares internacionales.

La OMS [2] publicó una base de datos sobre la contaminación atmosférica en las principales ciudades del mundo, en función de dos variables: PM_{10} y $PM_{2,5}$. De los 91 países del estudio, Chile ocupa el lugar 65. Los países con los registros más bajos de contaminación PM_{10}, presentan niveles nacionales del orden de los 12 a 15 $\mu g/m^3$ (por ejemplo, Estonia, Canadá, Australia, e Irlanda). Los países con los mayores niveles de contaminación incluyen India, Kuwait, Egipto, y Arabia Saudita con niveles superiores a los 100 $\mu g/m^3$. Chile presenta un nivel de 62 $\mu g/m^3$, un índice superior al promedio de los registros del estudio de la OMS de 57 $\mu g/m^3$.

Con respecto a la contaminación por $PM_{2,5}$, de los 38 países para los cuales existen información en la base de datos de la OMS, Chile ocupa el lugar 31, con un índice de $PM_{2,5}$ de 28,9. El valor promedio para este tipo de contaminante es 22 $\mu g/m^3$.

[1] Con fecha 18 de Febrero del 2013, publicado en: http://www.who.int/mediacentre/factsheets/fs313/es/

[2] Contaminación atmosférica en las ciudades del mundo. Congreso nacional de Chile. Disponible en: http://siit2.bcn.cl/actualidad-territorial/contaminacion-atmosferica-en-las-ciudades-del-mundo#_ftn2

En Chile en el año 2012 se modifica la norma de calidad del aire para el material particulado fino ($PM_{2,5}$). Esta modificación entró en vigencia el 1° de enero del 2012. Esta modificación implica cambios en el monitoreo de la calidad del aire tanto en el transporte urbano tales como autos, buses, camiones y así también las normas que deben cumplir los proyectos que tramita el servicio de evaluación de impacto ambiental. Con la modificación la norma[3] queda definida como sigue:

Cuadro 2.5: Niveles del $PM_{2,5}$ (leyChile)

	$PM_{2,5}$
Índice de media anual	$20\mu g/m^3$
Índice de media 24hrs.	$50\mu g/m^3$

2.2.5. Estación de monitoreo en Santiago de Chile

Santiago de Chile cuenta con una red de estaciones de monitoreo de la calidad de aire lo que nos permite contar con datos para realizar el análisis. Se han considerado registros existentes en el año 2012 - 2013 y 2014 de $PM_{2,5}$, temperatura, humedad relativa y velocidad del viento.

Figura 2.2.1: Estación de monitoreo, Santiago de Chile

[3]Disponible en: http://www.leychile.cl/Consulta/listaMasSolicitadasxmat?agr=1021&sub=514&tipCat=1

Capítulo 3

Conceptos de Geoestadística

3.1. Introducción

Supongamos que queremos estudiar la contaminación en la ciudad de Santiago de Chile y para ello contamos con datos que han sido registrados en estaciones de monitoreo durante algunos años. Ahora nos preguntamos ¿Cómo analizar estos datos? ¿Podemos pensar en una situación de laboratorio?¿Qué sugiere este tipo de problema?

Respecto al problema la literatura nos indica que debemos considerar la variabilidad espacial de los datos. Se ha desarrollado desde aproximadamente los años 30 una área de la estadística que resulta muy oportuna al evaluar las preguntas antes planteadas, *la geoestadística*. Esta área es relativamente nueva en la estadística y sus primeros trabajos fueron realizados por Fisher en el campo de la agricultura.

El desafío primordial de la estadística espacial (geoestadística) es modelar la variabilidad espacial. En los estudios de medio ambiente no es posible un diseño de experimento, ya que comúnmente nos enfrentamos a problemas asociados a las observaciones (frecuentemente la única información disponible, además es común que existan muchos datos faltantes, ya sea por problemas en los equipos o situaciones medio ambientales).

Los problemas medioambientales tiene asociados datos de diversa naturaleza, pudiendo ser continuos o discretos, ser agregados u observaciones individuales en determinados puntos, las localizaciones pueden ser regular o irregular. En los datos espaciales se pueden considerar:

- Observaciones de un fenómeno continuo en el espacio.
- Datos de una red de localizaciones
- Eventos que ocurren en el espacio proporcionando un conjunto aleatorio de puntos llamados patrón puntual.

Dependiendo de la naturaleza de los datos se da origen a las diferentes modelizaciones y análisis estadístico implicado. La dependencia espacial es la característica de los datos espaciales, y los modelos buscan recoger la estructura de interrelaciones.

3.2. Definición de geoestadística y campo de estudio

Es en el estudio de la minería que la geoestadística tiene su génesis. El estudio de la minería tiene diversas técnicas y métodos en la búsqueda de minerales útiles y la determinación de su concentración en determinadas zonas geográficas. Por ello se deben desarrollar diversas técnicas para estimar valores desconocidos a partir de los conocidos.

Uno de los pioneros de la geoestadística es el ingeniero en minas D. G. Krige, siendo George Matheron quien posteriormente consolida la geoestadística como disciplina de estudio de los datos espaciales.

La Geoestadística se centra en el análisis y el modelamiento de variables espaciales, en adelante variable regionalizada.

3.3. Conceptos básicos de geoestadística

3.3.1. Variable regionalizada

Cuando se habla de datos espaciales o geoestadísticos, asumiremos que estamos hablando de variables georeferenciadas, es decir tenemos mediciones en lugares determinados por un sistema de referencia. En nuestro problema, las estaciones de monitoreo registran las mediciones de las variables de interés, a su vez cada estación está ubicada en un sistema de referencia que será longitud, latitud y altura.

Formalmente [Cressie (1993)] define un proceso espacial en d dimensiones como: $\{Y(s) : s \in R^d\}$. Aquí Y denota la variable observada como por ejemplo concentración de $PM_{2,5}$. La localización en la cual Z es observado es s, que es un vector de coordenadas. Los tipos de datos espaciales son distinguidos según las características del dominio D.

El conjunto D es continuo y fijo. Por continuo entenderemos que $Y(s)$ puede ser observado en cualquier parte de D. De este modo entre localizaciones s_i y s_j, teóricamente se pueden encontrar un número infinito de otros puntos. Por fijo entenderemos que los puntos de D no son estocásticos. Es importante notar que la continuidad tiene relación con el dominio, no con el atributo que se está midiendo. Que el atributo Y sea continuo o discreto no tiene relación con que los datos sean geoestadísticos o no.

3.3.2. Estacionariedad

Cuando tratamos de hacer predicción, estamos pensando evidentemente en que el proceso estudiado tiene alguna regularidad o estabilidad susceptible de poder detectar, de lo contrario sería una tarea imposible. Frecuentemente se asume que el proceso espacial estudiado tiene una media $\mu(s) = E(Y(s))$ y que la varianza de $Y(s)$ existe para todo $s \in D$.

El proceso $Y(s)$ es Gaussiano, si para cualquier $n \geq 1$ y para cualquier conjunto de sitios $\{s_1, s_2, ..., s_n\}$, $Y = (Y(s_1), Y(s_2), ..., Y(s_n))^T$ tiene una distribución normal multivariante. Se dirá que el proceso es estrictamente estacionario si para cualquier $n \geq 1$ y para cualquier conjunto de sitios $\{s_1, s_2, ..., s_n\}$ y para cualquier $h \in \Re^d$, la distribución de $Y = (Y(s_1), Y(s_2), ..., Y(s_n))$ es la misma que $Y = (Y(s_1 + h), Y(s_2 + h), ..., Y(s_n + h))$. Este tipo de condición es muy fuerte y comúnmente es poco habitual, pues establece que las distribuciones conjuntas permanezcan invariantes ante una traslación. Es decir:

$$F_{S_1+h,...,S_m+h}(y_1, ..., y_m) \equiv F_{s_1,...s_m}(y_1, ..., y_m)$$

Una condición menos exigente es la estacionariedad de segundo orden o estacionariedad débil, ello es que la esperanza sea constante, es decir:

$$E\left(Y(s)\right) = \mu,\ \forall s \in D$$

Quedando:

$$cov\left(Y(s_1), Y(s_2)\right) = C\left(s_1 - s_2\right),\ \forall s_1, s_2 \in D$$

Así la función de covarianza de un proceso estacionario se puede expresar en un vector de diferencias entre los puntos.

Cuando la esperanza de la variable no es la misma en todas las direcciones o cuando la covarianza o correlación dependen del sentido en que se determinan, no habrá estacionariedad. Si la correlación no depende de la dirección en las que esta se calcule se dice que el fenómeno es isotrópico, en el caso contrario se habla de anisotropía.

Debido a que en la práctica resulta una tarea muy compleja el detectar la estacionariedad, también se estudia la estacionariedad considerando la variabilidad de los incrementos del proceso. En este caso se trabaja sólo con la hipótesis de que $[Y(s+h) - Y(s)]$ sean estacionarios, esto es:

$$E\left[Y(s+h) - Y(s)\right] = 0$$

$$Var\left(Y(s_1) - Y(s_2)\right) = 2\gamma(s_1 - s_2) = 2\gamma(h) = 2\left[C(0) - C(h)\right],\ \forall s_1, s_2 \in D$$

Esta propiedad se verifica si la varianza de las diferencias entre las variables en dos puntos depende únicamente del vector que las separa. A esta propiedad se llama estacionariedad intrínseca y es una condición más débil que la estacionariedad de segundo orden y habitualmente se utiliza en la práctica.

3.3.3. Funciones de correlación espacial

En el análisis estructural se estudia la dependencia espacial mediante el semivariograma, el covariograma y correlograma.

3.3.3.1. Variograma y semivariograma

Anteriormente se estableció que la estacionariedad asumía que la varianza de los incrementos de la variable regionalizada era finita. A esta función se le denomina variograma. De esta forma tenemos que:

$$2\gamma(h) = E\left(Y(s+h) - Y(s)^2\right)$$

A $\gamma(h)$ se le conoce como semivariograma y caracteriza la dependencia espacial del proceso. La función de semivariograma es estimada por el método de los momentos [Wanckernagel (1995)], a través del semivariograma experimental que se calcula mediante la fórmula:

$$\overline{\gamma}(h) = \frac{\sum \left(Y(s+h) - Y(s)\right)^2}{2n} \tag{3.3.1}$$

donde $Y(s)$ es la variable medida en el sitio s, $Y(s+h)$ es otro valor muestral medido a una distancia h del anterior y n es el número de parejas que se encuentran separadas por dicha distancia.

La función de semivariograma se calcula para varias distancias h. En la práctica comúnmente se consideran intervalos de distancias $\{[0,h], (h,2h), (2h,3h), ...\}$ y el semivariograma experimental corresponde a una distancia promedio entre parejas de sitios dentro de cada intervalo y no a una distancia total h específica.

Algunos parámetros del semivariograma son:

⊣ Efecto pepita c_0

Se refiere a que el variograma no tiende a 0 al acercarse al origen. Esto puede ser debido a un error de medida o la variación a muy pequeña escala. En algunas ocaciones puede indicar que parte de la estructura espacial se concentra a distancias inferiores a las observadas.

$$\lim_{h \to 0} \gamma(h) = c_0 > 0$$

⊣ Meseta c_s

El semivariograma crece con la distancia, delatando que en puntos cercanos el proceso es similar, hasta llegar a un punto que se estabiliza, cota denominada meseta. La meseta puede ser o no finita.

Los semivariogramas que tiene meseta finita cumplen con la hipótesis de estacionariedad extrínseca, mientras que cuando es infinita el semivariograma define un fenómeno natural que cumple solamente con la hipótesis de estacionariedad intrínseca.

$$\lim_{h\to\infty} \gamma(h) = c_s$$

⊣ Rango

Es la distancia h_s a la que se alcanza la meseta. Es decir corresponde a la distancia a partir de la cual dos observaciones son independientes.

El rango se denomina zona de influencia. Mientras más pequeño sea el rango más cerca se está del modelo de independencia espacial.

Para interpretar el semivariograma experimental se admite que a menor distancia entre los sitios entonces mayor correlación espacial. Por ello en presencia de autocorrelación se espera que para h pequeños el semivariograma experimental tenga magnitudes menores a las que este toma cuando las distancias de h se incrementan.

3.3.3.2. Covarianza y correlograma

La función de covarianza muestral entre parejas de observaciones que se encuentran a una distancia h se calcula, empleando la expresión clásica para la covarianza muestral:

$$C(h) = Cov\left(Y(s+h) - Y(s)\right) = \frac{\sum_{i=1}^{n}\left(Y(s+h)Y(s)\right) - m^2}{n}$$

donde m representa el valor promedio en todo punto de la región de estudio y n es el número de parejas de puntos que se encuentran a una distancia h. Si se asume estacionariedad y estimamos

la varianza de la variable regionalizada mediante la varianza muestral, se tiene que el correlograma muestral está dado por:

$$r(h) = \frac{Cov\left(Y(s+h) - Y(s)\right)}{S_{s+h} \cdot S_s} = \frac{C(h)}{S_s^2} = \frac{C(h)}{C(0)}$$

Bajo al supuesto de estacionariedad, se puede usar el semivariograma, el covariograma o correlograma para determinar la dependencia espacial de los datos. En la práctica comúnmente se utiliza el semivariograma .

3.3.3.3. Isotropía

Si el semivariograma $\gamma(h)$ depende del vector de separación sólo a través de su longitud $\|h\|$, entonces se entiende que el proceso es isotrópico. Así para un proceso isotrópico $\gamma(h)$, es una función del valor real de argumento univariado y se puede escribir como $\gamma\left(\|h\|\right)$. Si el proceso es intrínsecamente estacionario e isotrópico entonces el proceso es homogéneo.

Se utilizan diversos modelos isotrópicos de semivariograma entre ellos está el lineal, el esférico, exponencial, cuadráticos, racional, ondulado, potencial y gaussiano [Banerjee et al (2004)]. Estos modelos representan una serie de procesos espaciales.

3.3.4. Covarianza y variogramas

Consideremos que X tiene un semivariograma, entonces tenemos que:

⊣ Efecto pepita puro: Si $\exists\, \sigma^2 > 0$ tal que:

$$\gamma_x(h) = \begin{cases} \sigma^2 & si \qquad h \neq 0 \\ 0 & si \qquad h = 0 \end{cases}$$

⊣ Exponencial con parámetro $a > 0$ y $\sigma^2 > 0$ si:

$$\gamma_x(h) = \sigma^2 \left(1 - \exp\left(-\frac{\parallel h \parallel}{a}\right)\right)$$

⊣ Esférico con parámetro $a > 0$ y $\sigma^2 > 0$ si:

$$\gamma_x(h) = \begin{cases} \sigma^2 \left(\frac{3}{2} \frac{\| h \|}{a} - \frac{1}{2} \left(\frac{\| h \|}{a} \right)^3 \right), & si \quad \| h \| \leq a \\ \sigma^2 & si \quad \| h \| > a \end{cases}$$

⊣ Exponencial generalizado con parámetro $a > 0$; $\sigma^2 > 0$ y $0 < \alpha \leq 2$ si:

$$\gamma_x(h) = \sigma^2 \left(1 - \exp \left(- \frac{\| h \|}{a} \right)^{\alpha} \right)$$

si $\alpha = 2$ se llamará Gaussiano con parámetro $a > 0$ y $\sigma^2 > 0$.

⊣ Matérn

Se dice que X tiene semivariograma Matérn con parámetros $v > -1, a > 0$ y $\sigma^2 > 0$ si:

$$\gamma_x(h) = \sigma^2 \left(1 - \frac{2^{1-v}}{\Gamma(v)} \left(\frac{\| h \|}{a} \right)^v K_v \left(\frac{\| h \|}{a} \right) \right)$$

Nota: Sea $v > -1$ real. Se llama función de Bessel modificada de segundo clase con parámetro v a $K_v : [0, +\infty] \to \mathbb{R}$ dada por:

$$K_v(z) = \frac{\Gamma\left(v + \frac{1}{2}\right) (2z)^v}{\sqrt{\pi}} \int_0^{+\infty} \frac{\cos(t)}{\left(t^2 + z^2\right)^{v + \frac{1}{2}}} dt; \; z \geq 0$$

3.3.5. Función de covarianza

Con el fin de especificar un proceso estacionario se debe proporcionar una función de covarianza válida, esto es $c\,(h) \equiv cov\,(Y(s), Y(s+h))$ tal que en cualquier conjunto finito de sitios $s_1, s_2, ..., s_n$ y para cualquier $a_1, a_2, .., a_n$

$$Var \left[\sum_i a_i Y(s_i) \right] = \sum_{i,j} a_i \cdot a_j Cov\,(Y(s_i, Y(s_j))) = \sum_{i,j} a_i a_j c(s_i - s_j) \geq 0$$

Note que la desigualdad es estricta si no todos los a_i son 0. Necesitamos que $c(h)$ sea una función definida positiva, verificar esta condición no es trivial, pero el teorema de Bochner proporciona una condición suficiente y necesaria para que $c(h)$ lo sea. Este teorema es aplicado para h en el espacio d-dimensional euclidiano.

El teorema de Bochner establece que $c(h)$ es definida positiva si y sólo si:

$$c(h) = \int cos\left(w^T h\right) G\left(dw\right)$$

donde G es acotada, positiva, simétrica alrededor de 0 medida en $\Re^d$. Entonces $c(0) = \int Gd(w)$ se convierte es una constante normalizada y $\frac{G(dw)}{c(0)}$ es referida como la distribución espectral que induce a $c(h)$.

Por otro lado, si $G(dw)$ tiene una densidad con respecto a la medida de Lebesgue, es decir, $G(dw) = g(w)dw$, entonces $\frac{g(w)}{c(0)}$ es referida como la densidad espectral.

Debido a que $e^{iW^T h} = cos(w^T h) + isen(w^T h)$, tenemos que $c(h) = \int e^{iW^T h} G(dw)$. El término imaginario desaparece debido a la simetría de G alrededor de 0. Por otro lado $c(h)$ es una función válida si y sólo si es la función característica de una variable aleatoria simétrica d-dimensional (variable aleatoria con distribución simétrica). Notar que si G no se asume simétrica en 0, $c(h) = \int e^{iW^T h} G(dw)$ todavía proporciona una función de covarianza válida (definida positiva), pero ahora para un proceso aleatorio de valores complejos de $\Re^d$.

Considere que la transformada de Fourier para $c(h)$ es:

$$\hat{c}(w) = \int c(h) = \int e^{-iW^T h} c(h) dh \tag{3.3.2}$$

Al considerar y aplicar la fórmula inversa, $c(h) = (2\pi)^{-2} \int e^{iW^T h} \hat{c}(w) dw$ y se tiene que $(2\pi)^{-d} \hat{c}(w)/c(0) = g(w)$, la densidad espectral. El cálculo de 3.3.2 no es posible excepto en casos especiales. La relación entre $c(h)$ y $g(w)$ permite examinar los procesos espaciales en el dominio espectral en lugar del dominio observacional.

En lugar de buscar las funciones de correlación isotópicas que son válidas en todas las dimensiones, se puede buscar todas las funciones de correlación isotrópicas válidas en una dimensión particular d. [Matérn (1986)] proporciona un resultado general. Sea $c\left(\|h\|\right)$ de la forma:

$$c\left(\|h\|\right) = \int_0^\infty \left(\frac{2}{w\|h\|}\right)^\alpha \Gamma\left(v+1\right) J_v\left(w\|h\|\right) G(dw) \tag{3.3.3}$$

donde G es no decreciente e integrales en $\Re^+$, J_v es la función de Bessel de orden v y $v = \frac{d-2}{2}$ ofrece todas las funciones de correlación isotrópicas válidas en $\Re^d$.

El parámetro v en la clase Matérn es un parámetro de suavizamiento. En el espacio 2-dimensional, el valor entero más grande de v indica el número de veces en que el proceso será diferenciable. El uso de la función de covarianza Matérn como modelo permite que los datos informen sobre v, en definitiva podemos aprender sobre el suavizamiento del proceso, a pesar de observar el proceso sólo en un número finito de puntos.

La clase Matérn [Stein (1999)] es una herramienta general para la construcción de modelos espaciales. El cálculo de esta función requiere de una evaluación modificada de la función de Bessel.

3.4. Predicción espacial clásica

La predicción espacial clásica denominado kriging cuyo nombre procede del geólogo sudafricano D.G. Krige, cuyos trabajos en la predicción en la reserva de oro en la década de los 40 se consideran como los primeros trabajos dedicados a la interpolación espacial.

De la la teoría clásica de la decisión se conoce que si Y_0 es una cantidad aleatoria e Y_0^* es un predictor, entonces $L(Y_0, Y_0^*)$ representa la pérdida en que se incurre cuando se predice Y_0 con Y_0^* y el mejor predictor será el que minimice $E\{L(Y_0, Y_0^*) \mid Y\}$, con $Y = \{Y_1, Y_2, ..., Y_n\}$.

Si $Y(Y_0, Y_0^*) = E(Y_0 \mid Y)$, entonces para encontrar el predictor óptimo se requiere conocer la distribución de las $n+1$ variables aleatorias.

Un predictor para Y_0 basado en Y debe tener la forma $\sum \ell_i Y(s_i) + \delta_0$. Usando la pérdida del error cuadrático medio, el mejor predictor lineal será el que minimice:

$$E\left[Y(s_0) - \left(\sum \ell_i Y(s_i) + \delta_0\right)^2\right]$$

Para un proceso de media constante se tiene que $\sum \ell_i = 1$. En este caso se minimiza la expresión $E\left[Y(s_0) - (\sum \ell_i Y(s_i))^2\right] + \delta_0^2$ y δ_0 deberá ser 0.

Un variograma necesariamente debe satisfacer la condición definida negativa. De hecho para cualquier conjunto de localizaciones $s_1, s_2, ..., s_n$ y conjunto de $a_1, a_2, ..., a_n$ tales que $\sum a_i = 0$ y $\gamma(h)$ es válida, se cumple que $\sum_i \sum_j a_i a_j \gamma(s_i - s_j) \leq 0$.

$$\sum_i \sum_j a_i a_j \gamma(s_i - s_j) = -E\left[\sum a_i Y(s_i)\right]^2 \leq 0$$

Así si hacemos $a_0 = 1$ y $a_i = -\ell_i$, el predictor se convierte en $E\left[\sum_{i=0}^{n} a_i Y(s_i)\right]^2$ con $\sum a_i = 0$. Esta relación revela cómo históricamente el variograma surgió en el kriging.

Los ℓ óptimos pueden obtenerse resolviendo con multiplicadores de Lagrange la condición de optimización definida, los cuales serán funciones de $\gamma(h)$ [Cressie (1993)]. Con una estimación de $\gamma(h)$ se obtiene directamente el kriging ordinario.

El modelo para los datos observados viene dado por:

$$Y = \mu 1 + \epsilon, \quad \epsilon \sim N(0, \Sigma)$$

La estructura espacial de la covarianza sin considerar el efecto pepita, se define como:

$$\Sigma = \sigma^2 H(\phi), \quad H(\phi)_{ij} = \rho(\phi; d_{ij}) \tag{3.4.1}$$

donde $d_{ij} = \| s_i - s_j \|$ es la distancia entre s_i y s_j, y ρ es una función de correlación válida en $\Re^d$ [Banerjee et al (2004)].

Para un modelo con efecto pepita, se tendrá que Σ viene dada por:

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I \tag{3.4.2}$$

donde τ^2 es la varianza del efecto pepita.

Cuando se tiene el vector de covariables $x = (s(s_1), ..., x(s_n))^T$ y $x(s_0)$ disponibles para incorporarse en el análisis, el procedimiento anterior es denominado kriging universal. El modelo en este contexto queda de la siguiente forma:

$$Y = x\beta + \epsilon, \quad \epsilon \sim N(0, \Sigma) \tag{3.4.3}$$

donde Σ es definida como en 3.4.2 o 3.4.2 es decir con o sin efecto pepita.

El proceso de predicción se traduce en buscar la función $f(y)$ que minimice el error de predicción cuadrático medio, esto es:

$$E\left[Y(s_0) - f(y)^2 \mid y\right]$$

Se puede obtener que:

$$E\left[Y(s_0) - f(y)^2 \mid y\right] = E\left\{Y(s_0) - E\left[Y(s_0) \mid y\right]^2\right\} + \left\{E\left[Y(s_0) \mid y\right] - f(y)\right\}^2 \qquad (3.4.4)$$

En 3.4.4 la esperanza del término del producto cruzado es 0. Como el segundo término del lado derecho es no negativo, se obtiene que:

$$E\left[\left(Y(s_0) - f(y)\right)^2 \mid y\right] \geq E\left\{\left(Y(s_0) - E\left[Y(s_0) \mid y\right]\right)^2 \mid y\right\},$$

Para cualquier función $f(y)$. La igualdad se cumple si y sólo si $f(y) = E\left[Y(s_0) \mid y\right]$; de esta forma tenemos que el predictor $f(y)$ es justamente la media posterior de $Y(s_0)$ es decir, $f(y)$ minimiza el riesgo posterior.

Una vez identificada la forma del predictor veremos su estimación. Supongamos que los parámetros $\left(\beta, \sigma^2, \phi, \tau^2\right)$ son desconocidos y debemos estimarlos desde los datos. Modificamos la expresión de $f(y)$ de tal forma que:

$$\widehat{f(y)} = x_0^T + \hat{\gamma}^T \hat{\Sigma}^{-1}\left(y - X\hat{\beta}\right), \qquad (3.4.5)$$

donde $\hat{\gamma} = \left(\hat{\sigma}^2\rho(\hat{\phi}; d_{01}), ..., \hat{\sigma}^2\rho(\hat{\phi}; d_{0n})\right)^T$, $\hat{\beta} = \left(X^T\hat{\Sigma}^{-1}X\right)^{-1} X^T\hat{\Sigma}^{-1}y$, el estimador usual para β de mínimos cuadrados y $\hat{\Sigma} = \hat{\sigma}^2 H(\phi)$. Así $\widehat{f(y)}$ puede ser reescrita como $\lambda^T y$ con:

$$\lambda = \hat{\Sigma}^{-1}\hat{\gamma} + \hat{\Sigma}^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}\left(x_0 - X^T\hat{\Sigma}\hat{\gamma}\right)$$

3.5. Estadística Bayesiana

3.5.1. Conceptos básicos

El modelamiento Bayesiano se basa en el hecho de tratar a la distribución conjunta como una colección de variables aleatorias que se pueden descomponer en una serie de modelos adicionales.

Comúnmente en procesos complejos se suele escribir un modelo jerárquico en tres estados:

- Estado 1: Modelo para datos
- Estado 2: Modelo para el proceso dado parámetros del modelo
- Estado 3: Modelo para los parámetros

Los métodos Bayesianos permiten hacer estimaciones de forma natural en la modelización jerárquica. Bajo el enfoque Bayesiano la distribución a posteriori es obtenida a partir del teorema de Bayes. La dificultad del procedimiento tiene relación con la especificación de la distribución a priori de los parámetros involucrados.

3.5.2. Teorema de Bayes

Teorema 1. *Sea I un conjunto de índices numerables y sea $\{H_i\}_{i\in I}$ una partición sobre Ω. Si $A \in \mathscr{A}$ tal que $P(A) > 0$, entonces para todo $k \in I$,*

$$P\left(H_k \mid A\right) = \frac{P\left[A \mid H_k\right] \cdot P\left[H_k\right]}{P(A)} \tag{3.5.1}$$

$$= \frac{P\left[A \mid H_k\right] \cdot P\left[H_k\right]}{\sum_{i\in I} P\left[A \mid H_i\right] \cdot P\left[B_i\right]}$$

Una de las observaciones importantes del teorema de Bayes que se pueden mencionar ocurre cuando la partición H representa todas las posibles condiciones que se excluyen mutuamente; por ejemplo, los estados de la naturaleza, las hipótesis, etc, que son lógicamente posibles.

Por otro lado, las $P(H_1), P(H_2), \ldots$ representan el conocimiento previo, la experiencia, o las creencias en relación a la $P(H_i)$. Al observar un evento A, éste usualmente modifica o altera

las probabilidades de $H_1, H_2, \ldots$ .De aquí que $P(H_i)$ y $P(H_i \mid A)$ son llamadas probabilidades apriori y aposteriori de H_i, respectivamente.

3.5.3. Inferencia Bayesiana

En estadística clásica tenemos información contenida en la muestra y desde esa información estimamos parámetros. En estadística Bayesiana podemos contar con información previa acerca de los parámetros a través de su distribución de probabilidades. Esa información previa es utilizada con la información contenida en la muestra y usando el teorema de Bayes se puede obtener la distribución a posteriori de los parámetros, de la siguiente manera:

$$\pi\left(\theta \mid y\right) = \frac{\pi(\theta) L\left(y;\theta\right)}{\int_{\Theta} \pi(\theta) L\left(y;\theta\right) d\theta} \tag{3.5.2}$$

donde θ es el parámetro desconocido del modelo, es $\pi(\theta)$ es la distribución a priori para θ. $L\left(y;\theta\right)$ es la función de verosimilitud de θ y $\pi\left(\theta \mid y\right)$ es la distribución a posteriori de θ.

3.5.4. Distribución a priori

La información previa del parámetro θ puede ser incluida en el modelo a través de la distribución de probabilidades. Esta distribución se denomina a priori y se denota por $\pi(\theta)$. En esta distribución se debe incorporar toda la información que se tenga sobre los parámetros (teniendo presente valores que pueden asumir los parámetros). Si no hay información previa, se puede seguir utilizando apriori para dar la misma probabilidad para todos los valores de los parámetros posibles, en este caso, entonces tendremos una previa no informativa.

3.5.5. Función de verosimilitud

La función de verosimilitud es una función de los parámetros del modelo que agrega la información contenida en la muestra. Si tenemos observaciones independientes e idénticamente distribuidas de una distribución $f\left(y;\theta\right)$, donde Y es un vector de observaciones $(y_1, y_2, \ldots, y_n)$.

La función de verosimilitud está dada por:

$$L\left(y;\theta\right)=f\left(y_i;\theta\right)^n$$

3.5.6. Distribución a posteriori

Esta distribución es de especial interés para la estadística Bayesiana. Es una combinación entre la distribución apriori con la función de verosimilitud de los datos y describe la incertidumbre sobre el parámetro θ que se observa luego de la muestra. Cuando es posible encontrar esta distribución podemos hacer estimaciones puntuales como la media, mediana o moda. También es posible definir intervalos de confianza.

En la ecuación 3.5.2 el denominador es constante con respecto a θ. Entonces podemos decir que $\pi\left(\theta \mid y\right) \propto \pi\left(\theta\right) L\left(y;\theta\right)$.

3.6. Campos Gaussianos y Campos Aleatorios de Markov Gaussianos

Los campos Gaussianos (denotado por GF del inglés Gaussian Fields) tienen un importante rol en estadística espacial y constituye un componente en los modelos jerárquicos espaciales actuales.

Consideremos un dominio $D \in \Re^d$ con coordenadas $s \in D$. Entonces $s(x)$ es un GF continuamente indexado, si toda la colección finita $\{x(s_i)\}$ tiene conjuntamente distribución Gaussiana. En la mayoría de los casos, el GF se caracteriza con la media $\mu(\cdot)$ y una función de covarianza $C(\cdot)$, de modo que $\mu = (\mu(s_i))$ y la matriz de covarianza es $\Sigma = (C(s_i, s_j))$.

Frecuentemente la función de covarianza es sólo una función de posición relativa de dos localizaciones, en cuyo caso se dice que es estacionaria e isotrópica. Considerando que una matriz de covarianza es definida positiva, la función de covarianza debe ser una función positiva, debido a ello es complicado definir una forma cerrada para la función de covarianza. El teorema de Bochner permite caracterizar a todas las funciones continuas definidas positivas en $\Re^d$.

Un campo aleatorio de Markov Gaussiano (lo denotaremos por GMRF del inglés Gaussian Markov Random Fields) x, en un campo Gaussiano indexado discretamente, donde las condiciones completas $\pi(x_i \mid x_{i-1})$ con $i = 1, ..., n$ dependen solamente de un conjunto de vecinos ∂_i para las cada localización i (si $(i \in \partial_j)$ entonces $j \in \partial_i$). Considere que la notación $x = (x_1, ..., x_n)$ con $x \sim N(\mu, Q^{-1})$ se refiere a un GMRF n-dimensional con media μ y matriz de precisión simétrica y definida positiva Q (inversa de la matriz de covarianza).

3.7. Enfoque SPDE

Sea $x(s) = \{x(s), s\} \in D \subseteq \Re^2$ un campo Matérn, es decir, un GF estacionario de segundo e isotrópico con función de covarianza Matérn:

$$C(h) = \frac{1}{\Gamma(v)^{2v-1}}(kv)^v K_v(kh)$$

donde K_v es la función de Bessel de segundo tipo modificada .

La covarianza Matérn establece una relación entre el campo Gaussiano y la función de covarianza Matérn como una solución de ecuaciones diferenciales parciales estocásticas de la siguiente forma:

$$\left(k^2 - \triangle\right)^{\alpha/2} x(u) = W(u),\ u \in \Re^d,\ \alpha = v + \frac{d}{2},\ k > 0,\ v > 0$$

donde $\left(k^2 - \Delta\right)^{\frac{\alpha}{2}}$ es un operador pseudo-diferencial definido a través de sus propiedades espectrales [Whittle (1963)]. W es espacial Gaussiano, un ruido blanco con varianza unitaria; Δ es el Laplaciano:

$$\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$$

y de varianza marginal:

$$\sigma^2 = \frac{\Gamma(v)}{\Gamma(v + d/2)(4\pi)^{d/2} k^{2v}}$$

Capítulo 4

Modelo espacio temporal

4.1. Modelo espacio temporal.

Aquí $y(s_i, t)$ corresponde a las realizaciones de un proceso espacio temporal $Y(\cdot, \cdot)$ en las estaciones $i = 1, 2, ..., d$ localizadas en los sitios s_i en los tiempos $t = 1, ..., T$.

Para el estudio de los datos consideramos el siguiente modelo propuesto por [Cameletti et al (2012)], a saber:

$$y(s_i t) = z(s_i, t)\beta + \xi(s_i, t) + \epsilon(s_i, t) \tag{4.1.1}$$

donde

1. $z(s_i, t) = (z_1(s_i, t), ..., z_p(s_i, t))$ denota el vector de las p covariables para los sitios s_i en el tiempo t;

2. $\beta = (\beta_1, ..., \beta_p)^T$ son los parámetros;

3. $\epsilon(s_i, t) \sim N\left(0, \sigma_\epsilon^2\right)$ es un error de medición;

4. $\xi(s_i, t)$ es la realización del proceso de estado, es decir el verdadero valor no observado de la contaminación. Supondremos que es un campo Gaussiano espacio-temporal que cambia en el tiempo y que tiene la estructura de un proceso autoregresivo AR(1);

$$\xi(s_i,t) = a\xi(s_i,t-1) + \omega(s_i,t) \tag{4.1.2}$$

con $t = 2,...,T$ donde $\mid a \mid < 1$ y $\xi(s_i,1)$. Se deriva de la distribución estacionaria $N(0,\sigma^2_\omega/(1-a^2)$.

Por otra parte supondremos que $\omega(s_i,t)$ tiene una distribución gaussiana de media cero, se supone que es temporal independiente y se caracteriza por la función de covarianza espacio-temporal:

$$cov\left(\omega(s_i,t),\omega(s_j,t')\right) = \begin{cases} 0 & si \quad t \neq t' \\ \sigma^2_\omega C(h) & si \quad t = t', \end{cases}$$

para $i \neq j$. La función de correlación puramente espacial $C(h)$ depende solamente de la ubicación a través de la distancia euclidiana $h = \parallel s_i - s_j \parallel \in R$.

Por otro lado se asume que el proceso es de segundo orden fijo e isotrópico [Cressie (1993)] de ello se desprende que $var\left(\omega(s_i,t)\right) = \sigma^2_w$ para cada s_i y t.

Para la función de correlación $C(h)$ se utiliza la función de Matérn:

$$C(h) = \frac{1}{\Gamma(v)^{2v-1}}(kv)^v K_v(kh) \tag{4.1.3}$$

donde:

1. k_v denota la función de Bessel modificada de segunda clase y orden $v > 0$.
2. v mide el suavizamiento y su valor entero determina la diferencia en media cuadrática; y
3. $k > 0$ es un parámetro de escala relacionado con el rango ρ, es decir una distancia a la cual la correlación espacial tiende a cero. En particular se utiliza ρ de la definición empírica de la derivada $\rho = \dfrac{8v}{k}$ donde ρ corresponde a la distancia en que la correlación espacial está cerca de 0.1 para cada v.

Las ecuaciones 4.1.1 y 4.1.2 se pueden reescribir de la siguiente manera:

$$y_t = z_t\beta + \xi_t + \epsilon_t \qquad \epsilon \sim N(0,\sigma^2_\epsilon I_d) \tag{4.1.4}$$

y

$$\xi_t = a\xi_{t-1} + \omega_t \qquad \omega \sim N(0, \Sigma = \sigma_\omega^2 \tilde{\Sigma}) \tag{4.1.5}$$

donde:

1. I_d es la matriz identidad de dimensión d
2. $z_t = (z(s_1, t)', ..., z(s_d t)')'$
3. $\xi_t = (\xi(s_1 t), ..., \xi(s_d, t))'$

ξ_1 proviene de la función de estacionariedad del proceso AR(1) que se distribuye $N\left(0, {}^{\Sigma}/_{(1-a^2)}\right)$. $\tilde{\Sigma}$ es la matriz de correlaciones de dimensión d con elementos $C\left(\parallel s_i - s_j \parallel\right)$ donde $C(\cdot)$ es la función Matérn dada por 4.1.3 y está parametrizada por k y v.

Sea $\Theta = \left\{\beta, \sigma_\varepsilon^2, a, \sigma_\omega^2, \kappa\right\}$ el vector de parámetros a estimar. La distribución conjunta posteriori está dada por :

$$\pi(\theta, \xi \mid y) \propto \pi\left(y \mid \xi\right) \pi\left(\xi \mid \theta\right) \pi\theta \tag{4.1.6}$$

donde $\pi(\cdot)$ es la función de densidad de probabilidad, $y = \{y_i\}$ y $\xi = \{\xi_t\}$ con $t = 1, ..., T$ y $\pi(\theta) = \prod_{i=1}^{dim(\theta)} \pi(\theta_i)$. Así

$$\pi\left(\theta, \xi \mid y\right) \propto \left(\prod_{t=1}^{T} \pi(y_t \mid \xi, \theta)\right) \left(\prod_{t=2}^{T} \pi(\xi_t \mid \xi_{t-1}, \theta)\right) \tag{4.1.7}$$

[Cameletti et al (2012)]señala que a partir de las distribuciones gaussianas definidas en 4.1.4 y 4.1.5 se puede anotar la distribución conjunta:

$$\pi\left(\theta, \xi \mid y\right) \propto \left(\sigma_\varepsilon^2\right)^{-\frac{dT}{2}} exp\left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{t=1}^{T} \left(y_t - z\beta - \xi_t\right)' \left(y_t - z_t\beta - \xi_t\right)\right)$$

$$\times \left(\frac{\sigma_\omega^2}{1-a^2}\right)^{-\frac{d}{2}} \mid \widetilde{\Sigma} \mid^{-\frac{1}{2}} exp\left(-\frac{1-a^2}{2\sigma_\omega^2} \xi_1' \tilde{\Sigma}^{-1} \xi_1\right)$$

$$\times \left(\sigma_w^2\right)^{-\frac{d(T-1)}{2}} \mid \tilde{\Sigma} \mid^{-\frac{(T-1)}{2}} exp\left(-\frac{1}{2\sigma_w^2}\sum_{t=2}^{T}\left(\xi_t - a\xi_{t-1}\right)' \tilde{\Sigma}^{-1}\left(\xi_t - a\xi_{t-1}\right)\right)$$

$$\times \prod_{i=1}^{dim(\theta)} \pi(\theta_i)$$

Donde $\mid \tilde{\Sigma} \mid$ es el determinante de la matriz de varianza y covarianza Σ.

4.2. GMRF y SPDE

Un GMRF es un proceso espacial que modela la dependencia espacial de los datos observados en mallas regulares, retículos o regiones geográficas. La notación $x = (x_1, x_2, ..., x_n)'$ con $x \sim N\left(\mu, Q^{-1}\right)$ se refiere a un n-dimensional GMRF con media μ y matriz de precisión definida positiva Q.

La densidad de x está dada por:

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

La distribución completa de x_i $(i = 1, ..., n)$ depende sólo de unos pocos componentes de x_i, esto gracias a la propiedad Markoviana relacionada con la estructura de la vecindad. Si ∂_i constituye el conjunto de vecinos de cada unidad i,

$$\pi\left(x_i \mid x_{-i}\right) = \pi\left(x_i \mid x\partial_i\right)$$

donde x_{-i} denota los elementos de x pero sin x, esto equivale a decir que dada la estructura de vecindad ∂_i los termino x_i y $x_{\{-i,\partial_i\}}$ son independientes. [Rue y Held (2005)] usan la notación siguiente, para la relación de independencia:

$$x_i \perp x_{-i} = \pi(x_i \mid x\partial_i)$$

con $i = 1, ..., n$. El punto clave de esta relación de independencia tiene que ver con la matriz de precisión Q. Así para una pareja (i, j) cualquiera con $i \neq j$ se tiene que:

$$x_i \perp x_j \mid x_{-\{i,j\}} \Longleftrightarrow Q_{ij} = 0$$

Entonces el patrón de no-ceros de Q esta dado por la estructura de vecindad del proceso. Luego $Q_{ij} \neq 0$ si $j \in \{i, \partial_i\}$. Esto permite realizar rápidas factorizaciones de Q.

4.3. Enfoque SPDE

Sea $X(s) = \{x(s), s \in D \subseteq \mathbb{R}^2\}$ un campo Matérn, es decir un GF estacionario de segundo orden e isotrópico con función de covarianza dada en 4.1.3 y dependiente de los parámetros de escala y suavidad, κ y υ. Además supone observaciones del proceso s_i en d localizaciones $s_1, ..., s_d$.

El objetivo del enfoque SPDE es encontrar un GMRF, con estructura de vecindad y matriz dispersa de precisión Q que mejor represente el campo Matérn. Dada la representación es posible hacer inferencia usando el GMRF encontrado.

SPDE usa una representación finita para definir el campo Matérn como una combinación lineal de funciones base definidas en una triangulación en el dominio D. La triangulación consiste en dividir D en un conjunto de triángulos no interceptados unidos por al menos un borde y vértice en común. En primer lugar, los vértices de los triángulos iniciales son asignados en las localizaciones $s_1, ..., s_n$ y luego se agregan vértices adicionales en orden para obtener una triangulación.

La función base del campo Matérn $X(s)$ está dada por:

$$X(s) = \sum_{i=1}^{n} \psi_l(s)\omega_l \tag{4.3.1}$$

donde n es el número total de vértices, $\{\psi_l(s)\}$ son las funciones base y ω_l son los pesos con distribuciones Gaussianas. Las funciones $\{\psi_l(s)\}$ son seleccionadas para que sean trazos lineales en cada triángulo, es decir $\psi_l(s)$ es 1 en el vértice l y 0 en el otros vértices. La altura del triángulo (valor del campo espacial) está dado por w_l y los valores en el interior del triángulo son determinados por interpolación lineal.

La esencia del enfoque SPDE es la representación finita de 4.3.1 que establece el vínculo entre el GF $X(s)$ y el GMRF definido por los pesos Gaussianos ω_l a los cuales se les puede asignar una estructura markoviana, [Lindgren et al (2011)]

4.4. El modelo en R.

Para los tiempo $t = 1, ..., T$ el campo Matérn definido en 4.1.5 es representado mediante el GMRF $\tilde{\omega} \sim N\left(0, Q_s^{-1}\right)$, donde la matriz de precisión Q_s proviene de la representación SPDE [Lindgren et al (2011)]. La matriz Q_s no cambia en el tiempo debido a la hipótesis de independencia y su dimensión n está dado por el número de vértices del dominio de la triangulación. [Cameletti et al (2012)] escribe la ecuación 4.1.5 como :

$$\xi_t = a\xi_{t-1} + \tilde{\omega}_t \qquad\qquad \tilde{\omega}_t \sim N\left(0, Q_s^{-1}\right) \tag{4.4.1}$$

para $t = 1, ..., T$ y $\xi_1 \sim N\left(0, {}^{Q_s^{-1}}/_{(1-a^2)}\right)$. [Cameletti et al (2012)] señala que la distribución conjunta del GMRF $\xi = \left(\xi_1^{'}, ..., \xi_T^{'}\right)^{'}$ es:

$$\xi \sim N\left(0, Q^{-1}\right)$$

con $Q = Q_T \otimes Q_s$ donde:

$$Q_T = \begin{pmatrix} \sigma_\omega^2 & {}^{-a}/_{\sigma_\omega^2} & & & \\ {}^{-a}/_{\sigma_\omega^2} & {}^{(1+a^2)}/_{\sigma_\omega^2} & & & \\ & & ..... & & \\ & & & {}^{(1+a^2)}/_{\sigma_\omega^2} & {}^{-a}/_{\sigma_\omega^2} \\ & & & {}^{-a}/_{\sigma_\omega^2} & \sigma_\omega^2 \end{pmatrix}$$

es una matriz de precisión T dimensional del proceso AR(1) definido en 4.4.1. Se puede reescribir 4.1.4 :

$$y_t = z_t\beta + B\xi_t + \varepsilon_t, \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2 I_d\right) \tag{4.4.2}$$

donde la matriz B de dimensión $(n \times d)$ selecciona el valor del GMRF ξ_t para cada observación de y_t. En particular B es una matriz dispersa con un solo elemento para cada fila tal que:

$$y_i(s_i,t) = z(s_i,t)\beta + \sum_{j=1}^{n} B_{ij}\xi_t + \varepsilon(s_i,t)$$

donde $B_{ij} = 1$ si el vértice j del triángulo es de la localización s_i y 0 en otro lugar. [Cameletti et al (2012)]

El modelo jerárquico detallado en 4.4.1 y 4.4.2 es una clase modelo latente y puede estimarse usando INLA [Rue et al (2009)]. INLA es un enfoque computacional para la inferencia Bayesiana y es una alternativa a MCMC para conseguir los marginales posteriores aproximados para las variables latentes , así como para los hiperparámetros .

En el siguiente capitulo de la tesis se mostrará la implementación de INLA en R . En [R-INLA] podemos encontrar diversa documentación referente del paquete.

Capítulo 5

Aplicación del modelo a datos de la contaminación del aire en Santiago de Chile.

Para la aplicación se utiliza el software R, específicamente el paquete R-INLA. En primer lugar se realiza un análisis exploratorio de los datos, para luego definir modelo, estimar parámetros para finalmente realizar un mapa de predicción.

5.1. Región de estudio y variables

El estudio se realiza en la ciudad de Santiago de Chile, considerando datos del $PM_{2,5}$, temperatura (TEM), humedad relativa (HR) y velocidad del viento (VV), medidos en estaciones de monitoreo durante los años 2012-2013 y 2014. Los datos fueron facilitados por el Ministerio del Medio Ambiente.

5.1.1. Región de estudio desde GoogleMap

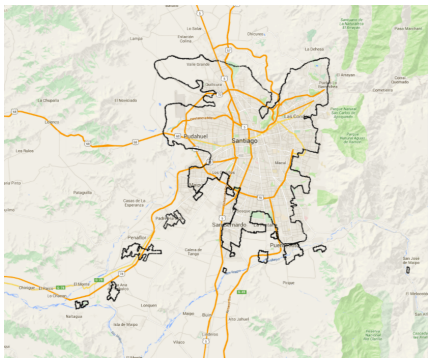
A continuación se muestra fotografía de Santiago de Chile, utilizando Google Map en R. Se ha trazado un contorno para identificar el área de estudio.

```
# carga de algunos puntos del contorno de la región de estudio.
```

```
Datos1 <- read.delim("cor3.txt",h=T,sep="\t",dec=",")
PM10DD<-data.frame(Datos1)
EST.XY=PM10DD[c("X1","X2")]
proj4string(PM10DD)<- CRS('+init=epsg:24879')

# Obtenemos el mapa desde GoogleMap
m<-plotGoogleMaps(PM10DD,filename='myMap1.htm')
```

Figura 5.1.1: Santiago de Chile, vista satelital.

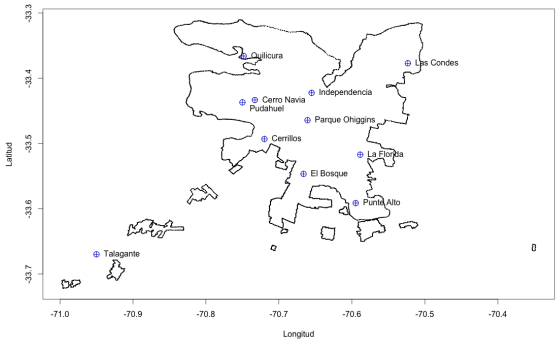


5.1.2. Contorno de la región de estudio con R

En la siguiente imagen se muestra el mapa de Santiago de Chile, pero ahora se ha trazado su contorno. Para ello basta tener disponible las coordenadas del perímetro de la región y luego se integra en R con el siguiente código.

```
# Mapa de Santiago de Chile
STGO<-read.table(arch1,header=TRUE)
plot(STGO,type="p",cex=.01,
col=8,xlab="Longitud",ylab="Latitud")
```

Figura 5.1.2: Contorno de Santiago de Chile con estaciones de monitoreo.



5.1.3. Estaciones de monitoreo.

Considere y_i: $PM_{2,5}$ medido en las estaciones de monitoreo i. Luego se debe considerar lo siguiente:

Cuadro 5.1: Estaciones de monitoreo (Santiago de Chile)

i	Notación secundaria	Estación
1	F	La Paz (Independencia)
2	L	La florida
3	M	Las condes
4	N	Santiago(P. O'Higgins)
5	O	Pudahuel
6	p	Cerrillos
7	Q	El Bosque
8	R	Cerro Navia
9	S	Punte Alto
10	T	Talagante
11	V	Quilicura

El registro del $PM_{2,5}$ se ha realizado cada hora durante el los años 2012-2013 y 2014. Para efectos de este estudio se considera un promedio diario de dicha concentración.

5.1.4. Base de datos del $PM_{2,5}$

La base de datos considera 11 estaciones de monitoreo y un registro cada hora desde el 01/01/2012 hasta el 31/12/2014. Para el trabajo se considera un promedio diario por estación, es decir una media cada 24 registros.

De esta manera todo el análisis posterior se realiza con promedio diario. El cálculo del promedio diario del $PM_{2,5}$ se realiza con el siguiente código

```
pm_2012<-read.delim("2012v.txt",h=T,sep="\t",dec=",")
pm_2013<-read.delim("2013v.txt",h=T,sep="\t",dec=",")
pm_2014<-read.delim("2014v.txt",h=T,sep="\t",dec=",")
Dat<-as.matrix(pm)
S<-dim(Dat) S prom<-c() for(i in 1:(S[1]/24)){
prom<-rbind(prom,colMeans(Dat[((i-1)*24):(i*24),6:S[2]],na.rm=T)) }
prom
```

Figura 5.1.3: Estructura de la base de datos.

Station.ID	FECHA	LONGITUD	LATITUD	TEM	HR	VV	PM2.5
1	1-1-2012	-70.65518	-33.42232	23.91208	52.93750	1.246667	25.33333
2	1-1-2012	-70.58859	-33.51716	23.13042	55.93042	1.239583	28.83333
4	1-1-2012	-70.66077	-33.46429	22.19625	57.54087	1.576250	26.33333
5	1-1-2012	-70.75000	-33.43700	22.99833	58.13000	2.323750	25.83333
7	1-1-2012	-70.66625	-33.54671	21.88542	58.89478	1.918750	25.54167
9	1-1-2012	-70.59478	-33.59138	21.10625	61.39583	2.217917	29.20833
10	1-1-2012	-70.95000	-33.67000	19.14042	69.95250	1.277917	22.29167
11	1-1-2012	-70.74823	-33.36587	23.14333	55.28000	2.182917	24.58333
1	2-1-2012	70.65518	-33.42232	26.34792	45.05750	1.278750	23.12500

5.2. Análisis exploratorio.

5.2.1. Resumen del $PM_{2,5}$ en las estaciones de monitoreo.

A continuación se resume información del $PM_{2,5}$ en cada estación de monitoreo según año.

Cuadro 5.3: Resumen del $PM_{2,5}$ por estación

	Año	F	L	M	N	O	P	Q	R	S	T	V
Promedio	2012	23	26	19	26	26	25	28	29	28	21	24
	2013	24	25	21	27	24	26	29	27	35	21	25
	2014	28	29	23	24	29	31	32	33	27	19	28
Desviación estándar	2012	11	11	8	14	19	14	15	21	11	14	13
	2013	11	10	7	12	15	13	15	17	15	12	11
	2014	16	18	13	12	24	20	22	29	15	12	19

Esto nos permite ya tener una idea sobre la variabilidad de los datos. En particular observar que la estación Cerro Navia (R) es la que presenta mayor promedio en año 2012 y en el año 2014. La estación Puente alto es la que presenta mayor promedio en el año 2013.

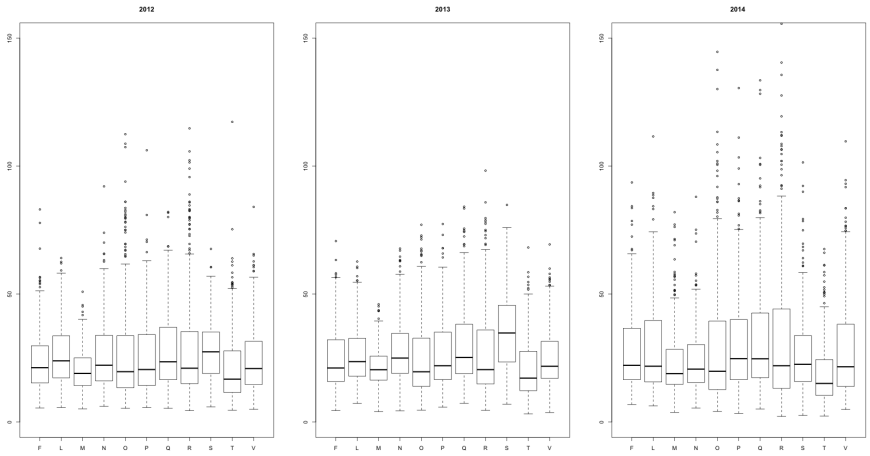
Las estaciones con menor promedio son Las Condes(M) en el año 2012, Las Condes (M) y Talagante en el año 2013 y Talagante (T) en el año 2014.

Las estaciones La Paz (F), La Florida(F), Las Condes(M), Cerrillos(P), El Bosque (Q), Talagante(T) y Quilicura (V) existe un alza cada año en las concentraciones promedios del $PM_{2,5}$.

5.2.2. Box plot del $PM_{2,5}$

La siguiente gráfica muestra una vista de las concentraciones del $PM_{2,5}$ por año.

Figura 5.2.1: box-plot del $PM_{2,5}$ por estación según año.



Aquí se puede observar que la estación Talagante tiene una mediana menor en cada año. En cambio es la estación Puente Alto (S) la que presenta mayor mediana en los años 2012 y 2013. En el año 2014 Cerrillos (P) y el Bosque (Q) tiene mayor mediana. Notar que el general existen valores extremos de las mediciones del $PM_{2,5}$ en la mayoría de las estaciones. En general se observa que los datos son levemente asimétricos.

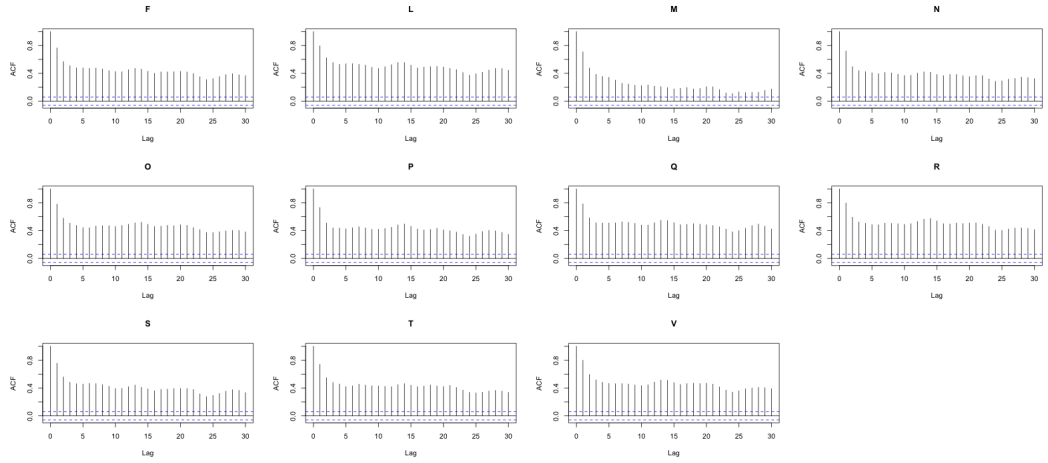
5.2.3. Serie de tiempo de los datos y autocorrelación.

Se muestra gráficas temporales de los datos, se ha realizado test para evaluar la autocorrelación simple de los datos (AFC) . Si dos o más barras de nuestro diagrama quedan fuera de las bandas de confianza se rechaza la hipótesis de independencia de los datos. Esta hipótesis se basa en que las mediciones son dependientes de sus anteriores en el tiempo. En nuestro caso se evidencia clara dependencia temporal de los datos.

```
# serie de tiempo por cada estación
ts_f<-ts(prom[,1])
ts_l<-ts(prom[,2])
ts_m<-ts(prom[,3])

# gráficas por estación
par(mfrow=c(2,5))
acf(ts_f,na.action = na.pass,main="F")
acf(ts_l,na.action = na.pass,main="L")
acf(ts_m,na.action = na.pass,main="M")
```

Figura 5.2.2: Autocorrelación temporal



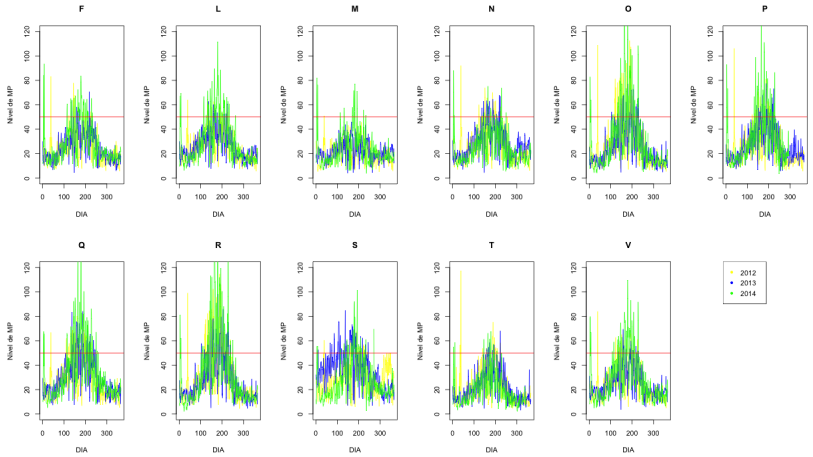
5.2.4. Comportamiento temporal del $PM_{2,5}$

En la siguiente gráfica se muestra la evolución temporal del $PM_{2,5}$ (promedio diario). Se ha trazado una recta en $y = 50$, correspondiente al límite diario del $PM_{2,5}$ según la legislación chilena.mPara originar la gráfica se utiliza el siguiente código.

```
# F
plot(prom[,1],type="l",ylim=c(0,120),col="yellow",xlab="DIA",ylab="Nivel de MP",main="F")
lines(prom[,12],type="l",col="blue")
lines(prom[,23],type="l",col="green")
abline(50,0,col="red")
# L
plot(prom[,2],type="l",ylim=c(0,120),col="yellow",xlab="DIA",ylab="Nivel de MP",main="L")
lines(prom[,13],type="l",col="blue")
```

```
lines(prom[,24],type="l",col="green")
abline(50,0,col="red")
```

Figura 5.2.3: Nivel del $PM_{2,5}$ por día según estación.



Se observa que los datos en cada año tienen un comportamiento muy similar. La norma se supera en los días centrales de cada año (al rededor del día 200).

A continuación se muestra una tabla que resume los días que se supera la norma por estación según año.

Cuadro 5.4: Número de días que se supera la norma por estación, según año.

Año	F	L	M	N	O	P	Q	R	S	T	V	Total de días
2012	13	13	1	33	41	25	43	57	17	18	24	285
2013	8	15	0	25	29	21	36	49	57	7	18	258
2014	46	57	17	16	58	49	69	71	39	12	54	488

Aquí se puede observar que en el año 2012 las estaciones de Santiago (N), Pudahuel (O) y Cerro Navia (R) son las que presentan más de 30 días en los cuales se supera la norma.

En el año 2013, El Bosque (Q), Cerro Navia (R) y Puente Alto (S) tienen más de 30 días en que se supera la norma y en el año 2014, La paz (F), La florida (L), Pudahuel (O), Cerrillos

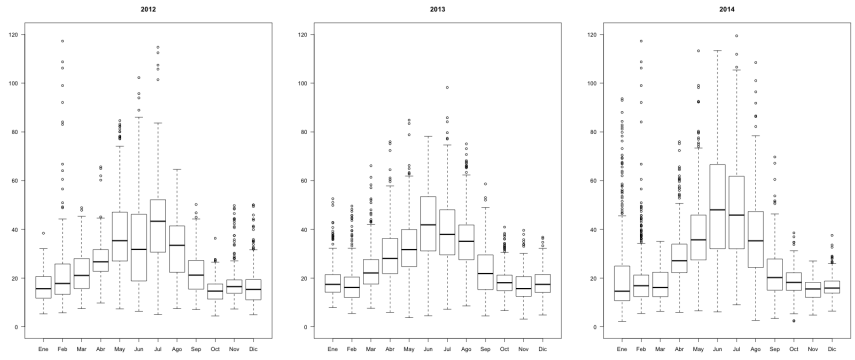
(P), El Bosque (Q), Cerro Navia (R), Puente Alto (S) y Quilicura (V) tienen más de 30 días en que se supera la norma.

Notar que existe un aumento importante del número de días que se supera la norma en el año 2014 comparado con los años anteriores.

5.2.5. $PM_{2,5}$ en cada mes

Aquí se muestra la concentración de $PM_{2,5}$ en cada mes. Notar que en los meses de invierno en donde se observan los mayores índices de la contaminación.

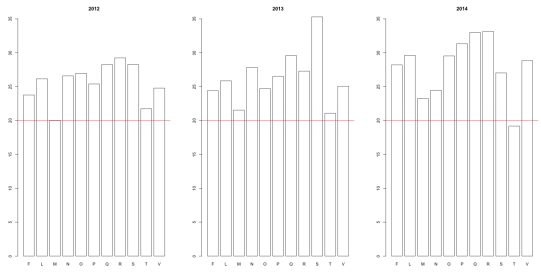
Figura 5.2.4: Nivel del $PM_{2,5}$ por mes según año.



5.2.6. Resumen anual

Podemos ver el promedio anual de la concentración del $PM_{2,5}$ por cada estación. Se ha trazado una recta en $y = 20$, que corresponde a la media anual según la legislación vigente en Chile. Se puede observar que todas las estaciones superan la norma anual salvo la estación Talagante (T) en el año 2014.

Figura 5.2.5: Nivel del $PM_{2,5}$ anual



5.2.7. Covariables

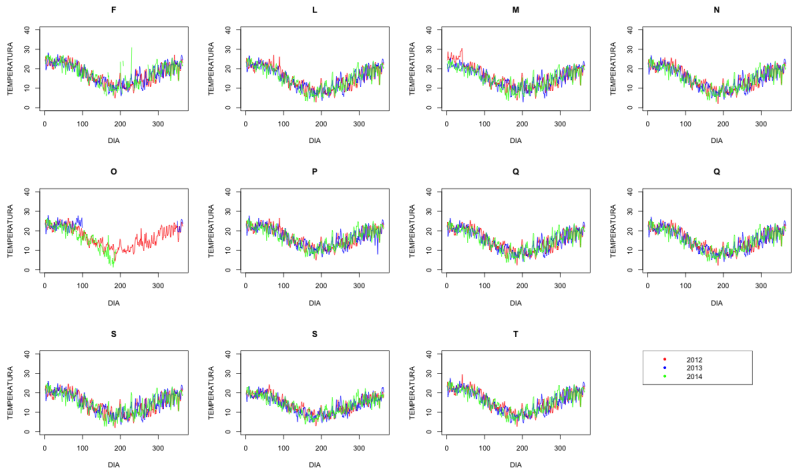
En el estudio se considera la temperatura, la humedad relativa y la velocidad del viento como covariables, además en el análisis se incorporan covariables geográficas que corresponden a la longitud y latitud.

x_j	Nombre de covariable	Notación secundaria
x_1	Temperatura	TEM
x_2	Humedad relativa	HR
x_3	Velocidad del viento	VV

5.2.8. Comportamiento temporal de la Temperatura.

En la gráfica se muestra el comportamiento de la temperatura por estación según año. Se puede observar que tienen un comportamiento muy similar en cada estación según el año. Además se observa que en todos los casos en los meses de invierno disminuye la temperatura.

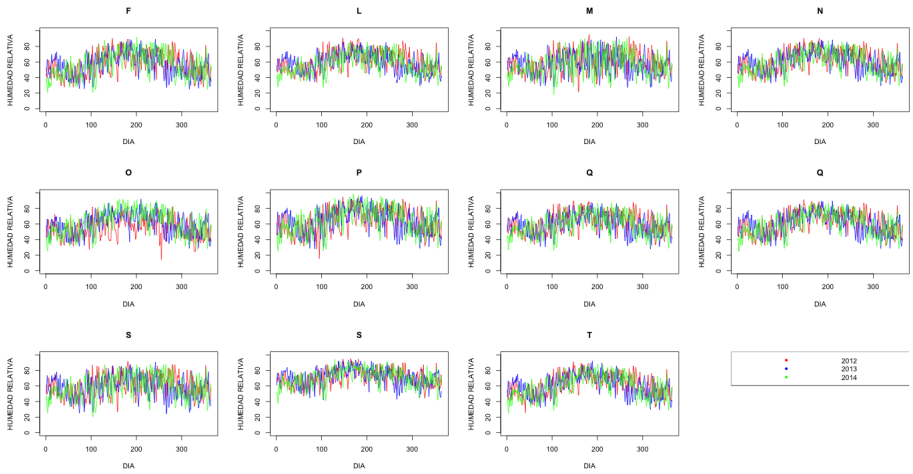
Figura 5.2.6: Temperatura por estación según año.



5.2.9. Comportamiento temporal de la Humedad Relativa

En la gráfica se muestra el comportamiento de la humedad relativa por estación según año. Se puede observar que tienen un comportamiento muy similar en cada estación según el año.

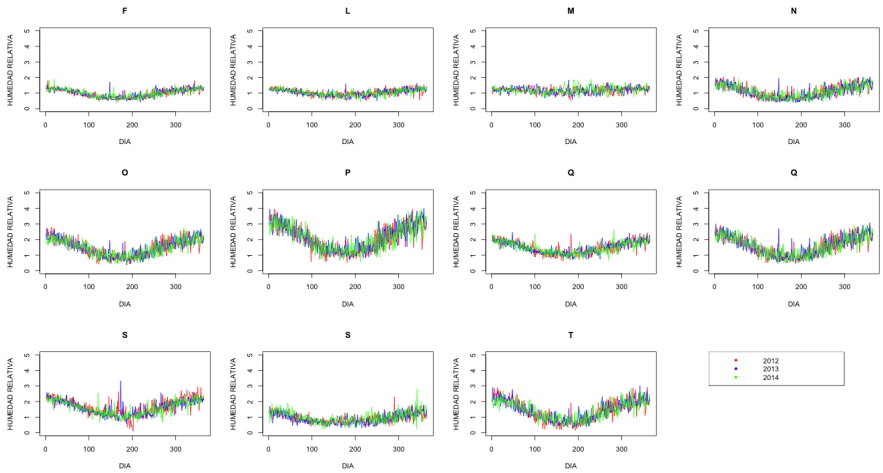
Figura 5.2.7: Humedad relativa por estación según año.



5.2.10. Comportamiento temporal de la velocidad del viento

En la gráfica se muestra el comportamiento de la velocidad del viento por estación según año. Se puede observar que tienen un comportamiento muy similar en cada estación según el año.

Figura 5.2.8: Velocidad del viento por estación según año.



5.2.11. Modelo lineal simple

Para analizar la contribución de las covariables temperatura, humedad relativa y velocidad del viento a explicar la concentración del $PM_{2,5}$ se define un modelo lineal. El modelo tiene la forma:

$$PM_{2,5} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Donde x_1, x_2 y x_3 es la temperatura, humedad relativa y velocidad del viento respectivamente.

En R se define de la siguiente forma $lm\,(PM_{2,5} \sim x_1 + x_2 + x_3)$ obteniendo la siguiente estimación de parámetros.

Cuadro 5.6: Coeficientes del modelo

Coeficiente	Estimado	Error estándar	p valor
Intercepto	67.61351	1.17366	<2e-16 ***
Temperatura	-1.19998	0.03699	<2e-16 ***
Humedad relativa	-0.22288	0.01259	<2e-16 ***
Velocidad del Viento	-6.69697	0.25462	<2e-16 ***

Aquí interesa ver que tanto la temperatura, la humedad relativa y la velocidad de viento contribuyen de manera significa a explicar el $PM_{2,5}$.

5.3. Resultados finales

5.3.1. Base de datos (DATOS)

De la base de datos original contiene registros de la temperatura, humedad relativa, velocidad del viento y del $PM_{2,5}$ medidos en las 11 estaciones de monitoreo de la ciudad de Santiago de Chile. Estos registros fueron realizados cada hora durante los años 2012-2013 y 2014.

De la base de dato original se ha considerado un promedio diario de la medición de cada variable. En adelante:

- DATOS: Contiene la información de 8 de las estaciones de monitoreo.
- DATOS_ VAL: Contiene la información de 3 de las estaciones consideradas estaciones de validación para el modelo a implementar.

A continuación se muestra la estructura general de la base de datos (9 primeros registros de DATOS)

Figura 5.3.1: Estructura de la base de datos.

Station.ID	FECHA	LONGITUD	LATITUD	TEM	HR	VV	PM2.5
1	1-1-2012	-70.65518	-33.42232	23.91208	52.93750	1.246667	25.33333
2	1-1-2012	-70.58859	-33.51716	23.13042	55.93042	1.239583	28.83333
4	1-1-2012	-70.66077	-33.46429	22.19625	57.54087	1.576250	26.33333
5	1-1-2012	-70.75000	-33.43700	22.99833	58.13000	2.323750	25.83333
7	1-1-2012	-70.66625	-33.54671	21.88542	58.89478	1.918750	25.54167
9	1-1-2012	-70.59478	-33.59138	21.10625	61.39583	2.217917	29.20833
10	1-1-2012	-70.95000	-33.67000	19.14042	69.95250	1.277917	22.29167
11	1-1-2012	-70.74823	-33.36587	23.14333	55.28000	2.182917	24.58333
1	2-1-2012	70.65518	-33.42232	26.34792	45.05750	1.278750	23.12500

5.3.2. Base de datos de validación (DATOS_VAL)

La estructura general de la base de datos de validación es la siguiente (6 primeros registros de DATOS_VAL).

Figura 5.3.2: Estructura de la base de datos de validación.

Station.ID	FECHA	LONGITUD	LATITUD	TEM	HR	VV	PM2.5
3	1-1-2012	-70.52331	-33.37697	24.25667	59.19750	1.129583	30.29167
6	1-1-2012	-70.71985	-33.49311	22.19833	61.91250	3.343750	24.33333
8	1-1-2012	-70.73290	-33.43332	22.00958	57.01091	2.487500	23.66667
3	1-1-2012	-70.52331	-33.37697	24.62786	60.81500	0.975000	27.57143
6	1-1-2012	-70.71985	-33.49311	25.14417	51.08542	3.375833	22.47826
8	1-1-2012	-70.73290	-33.43332	24.96083	48.48417	2.160833	20.83333

Cálculo de número de estaciones y largos de vectores.

```
N_ESTACIONES <- length(COORD_ESTACIONES$Station.ID)
N_ESTACIONES_VAL <- length(COORD_VALIDACION$Station.ID)
N_DATOS <- length(DATOS$Station.ID)
N_DIAS <- as.integer(N_DATOS/N_ESTACIONES)
```

Se estandariza variables además se incorpora columnas de $log(PM_{2,5})$ y vector de tiempos.

```
MEDIA_COVARIABLES = apply(DATOS[,3:7],2,mean,na.rm=T)
SD_COVARIABLES = apply(DATOS[,3:7],2,sd,na.rm=T)
```

```
DATOS[,3:7]=scale(DATOS[,3:7],MEDIA_COVARIABLES, SD_COVARIABLES)

# Se crea vector de tiempos
DATOS$logPM2.5 = log(DATOS$PM2.5)
DATOS_VALIDACION$logPM2.5 = log(DATOS_VALIDACION$PM2.5)
DATOS$time = rep(1:N_DIAS,each = N_ESTACIONES)
DATOS_VALIDACION$time = rep(1:N_DIAS,each = N_ESTACIONES_VAL)
```

De esta manera la base de datos queda ahora de la siguiente forma:

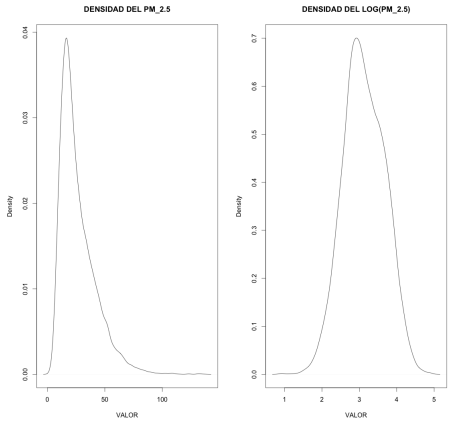
Figura 5.3.3: Estructura de la base de datos (estandarizada)

Station.ID	FECHA	LONGITUD	LATITUD	TEM	HR	VV	PM2.5	log(PM2_5)	TIME
1	1-1-2012	0.4260277	0.8564013	1.5483926	-0.50802195	-0.065985938	25.33333	3.232121	1
2	1-1-2012	1.0355691	-0.1649747	1.4036248	-0.29516868	-0.080024587	28.83333	3.361532	1
4	1-1-2012	0.3748787	0.4044069	1.2306133	-0.18063487	0.587224153	26.33333	3.270836	1
5	1-1-2012	-0.4418768	0.6983056	1.3791624	-0.13873649	2.068714547	25.83333	3.251666	1
7	1-1-2012	0.3247182	0.4832124	1.1730457	-0.08434591	1.266034133	25.54167	3.240311	1
9	1-1-2012	0.9789097	0.9642844	1.0287408	0.09352634	1.858960611	29.20833	3.374454	1
10	1-1-2012	-2.2725513	-1.8109798	0.6646604	0.70206799	-0.004050720	22.29167	3.104213	1
11	1-1-2012	-0.4256753	1.4643376	1.4060170	-0.34142567	1.789593168	24.58333	3.202069	1
1	2-1-2012	0.4260277	0.8564013	1.9995190	1.06843974	-0.002399115	23.12500	3.140914	2

5.3.3. Densidad del PM.

Se muestra en la gráfica la densidad del los datos del $PM_{2,5}$ y de los datos $log(PM2{,}5)$ mostrados anteriormente.

Figura 5.3.4: Densidad $PM_{2,5}$ y $log(PM_{2,5})$

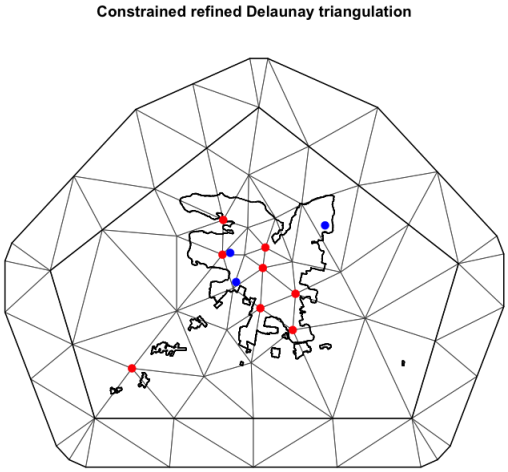


5.3.4. Triangulación de la región:

Usando INLA, se ha obtenido la triangulación de región, $inla.mesh.create.helper$ nos permite crear una malla con pequeños triángulos en el dominio de interés, y triángulos más grandes en la extensión. Se consideran los puntos iniciales correspondientes a las estaciones de monitoreo (8 estaciones). loc controla los puntos donde comienza la triangulación , en nuestro caso corresponden a las coordenadas de las estaciones de monitoreo (coordenada X (Longitud), CoordenadaY (Latitud)). $loc.domain$ permite definir un dominio que puede ser un polígono, con $inla.mesh.create.helper$ debemos proporcionar una malla convexa, en nuestro caso el dominio que corresponde al borde de la región (Coordenadas del borde de Santiago de Chile). Lo anterior se debe especificar obligatoriamente.

Además podemos especificar condiciones adicionales, *offset* permite controlar la distancia de extensión. Uno o dos valores, por una interna y una extensión exterior opcional. Si es negativo, interpretado como un factor en relación con el diámetro aproximado de datos (por defecto = -0.10). Se utiliza un conjunto convexo con 5 bordes a una distancia de 0.05 km. $max.edge$ controla la longitud del lado más grande permitido para el triángulo. La longitud máxima de lado se especifica a 0.3 km en el interior de la región, y 0.5 km en la extensión exterior. Dado que la extensión exterior no es de interés práctico, la resolución sólo tiene que ser tan fina como sea necesario por los valores numéricos. $min.angle$ permite controlar el ángulo interior permitido mínimo para los triángulos, 21 grados es el ángulo recomendado para el algoritmo. Se usa 26 grados para la región interior y 21 grados para la extensión exterior.

Figura 5.3.5: Triangulación con INLA.



El código utilizado es:

```
# TRIANGULACION DE LA REGION USANDO inla.mesh
MESH = inla.mesh.create.helper(points=cbind(COORD_ESTACIONES$X,
COORD_ESTACIONES$Y),
points.domain=BORDE,
n=5,
offset=c(0.05,0.1),
max.edge=c(0.3, 0.5),
min.angle=c(26, 21))
plot(MESH)
```

5.3.5. Definición del Modelo

En primer lugar se debe crear una matriz de covarianza espacial Matérn, para ello se usa $inla.spde2.matern(.)$ especificando el parámetro de suavizamiento $\alpha = 2$.

También se utiliza $inla.stack(\cdot)$ para construir las estructuras de datos necesarias, esto frecuentemente se utiliza para modelo complejos, en este caso es un modelo jerárquico.

Además se construye una matriz de observación que extrae los valores del campo espacio-temporal en los lugares de medición y el tiempo, puntos utilizados para la estimación de parámetros. El modelo completo utiliza una combinación de un modelo espacio-temporal latente y efectos de covarianza. Aquí se utiliza $inla.spde.make.index$, que genera vectores de los índices

de los componentes espaciales y temporales del modelo. Genera una lista con el campo y los campos $field.group$, en el que el primero contiene índices de vértices espaciales y el último contiene índices temporales.

Finalmente se define el modelo aditivo, especificando un modelo autoregresivo.

Se muestra parte del código usado:

```
spde = inla.spde2.matern(mesh=MESH, alpha=2)
 A.est =inla.spde.make.A(MESH,
loc=as.matrix(COORD_ESTACIONES[DATOS$Station.ID,
c("X","Y")]),
group=DATOS$time,
n.group=N_DIAS )
.
.
.
group=DATOS_VALIDACION$time,
n.group=N_DIAS)
.
.
.
# DEFINICION DEL MODELO
formula <- (logPM2.5 ~ -1 +
Intercept
+ UTMX
+ UTMY
+ TEM
+ HR
+ VV
+ f(field,
model=spde,
group=field.group,
control.group=list(model="ar1")))
```

5.3.6. Parámetros del Modelo

Considerando el modelo se han obtenido los siguientes resultados.

Cuadro 5.7: Parámetros del modelo

	mean	sd	q1	q2	q3
Intercepto	3.4856	0.6764	2.0652	3.4829	4.9167
Longitud	-0.8279	0.0263	-0.8279	-7.3661	-0.7770
Latitud	-0.3251	0.0230	-0.3703	-0.3251	-0.2800
Temperatura (TEM)	-0.7920	0.0386	-0.8678	-0.7921	0.7161
Humedad relativa (HR)	-0.3372	0.0293	-0.3948	-0.3372	-0.2797
Velocidad del viento (VV)	0.6505	0.0285	-0.5946	0.6505	0.7065

La salida de INLA proporciona la media de los parámetros además de los cuartiles. El valor del intercepto se de $3{,}48$ es decir el valor real es $exp(3{,}48)$ obtenemos el valor de $32{,}45$, lo que se interpreta como una potente explicación del nivel del $PM_{2{,}5}$ por las covariables. Además un muy buen antecedente es que la desviaciones estándar es muy pequeñas.

La estimación de los parámetros se ha obtenido con el siguiente código:

```
if (!exists("do.remote") || do.remote) {
result =
inla(formula,
data=inla.stack.data(stack, spde=spde),
family="gaussian",
control.predictor=list(A=inla.stack.A(stack), compute=TRUE),
control.compute=list(cpo=FALSE),
control.inla = list(reordering = "metis"),
keep=FALSE, verbose=TRUE,
inla.call="remote", num.threads=12)
}
.
.
.
```

También se han obtenido los hiperparámetros del modelo. El enfoque SPDE utiliza varianza local τ^2 de manera que $\sigma^2 = \dfrac{1}{2\pi\kappa^2\tau^2}$. INLA trabaja con $log(\tau^2)$ y $log(\kappa)$. Se puede obtener σ^2 con algunos cálculos. La siguiente tabla muestra la salida de INLA.

Cuadro 5.8: Hiperparámetros

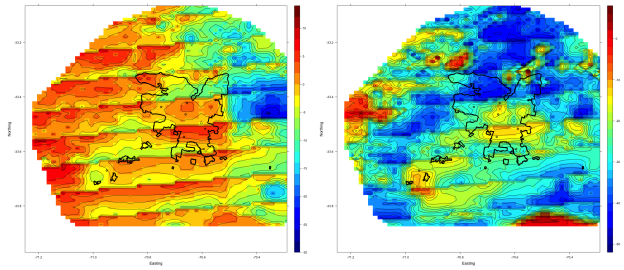
	mean	sd	q1	q2	q3
σ_ε^2	0.3316	0.0051	0.3219	0.3314	0.3421
σ_w^2	-3.3731	0.2341	-3.8395	-3.3703	-2.9197
ρ	2.1260	0.2409	1.6359	2.1335	2.5819
a	0.9985	0.0008	1.0000	0.9986	0.9997

5.3.7. Mapa de estimación del $PM_{2,5}$

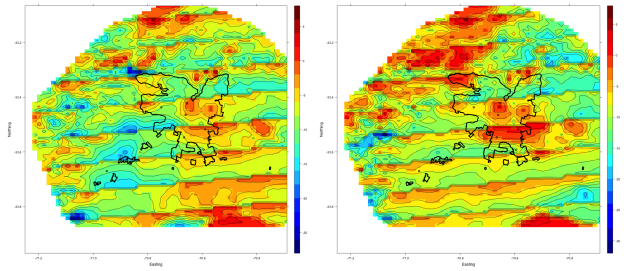
Para la predicción espacial se ha definido un área de 150km la que se han subdividido formando un grilla (168×72). De esta forma se obtiene un mapa para un día determinado, la idea básica es construir un arreglo de datos sobre las grillas que sea de la forma $(x, y, día)$ es decir cubos de datos dónde en la profundidad se tienen matrices para cada día que en este caso va desde 1 a 1095. En definitiva se define un modelo lineal simple para $campo\ latente + covariable\ geográfica$.

Cuadro 5.9: Ejemplo de Mapas.

(a) 01 de Julio 2012 (b) 02 Febrero de 2012



(c) 20 de Noviembre 2012 (d) 10 de Abril de 2012

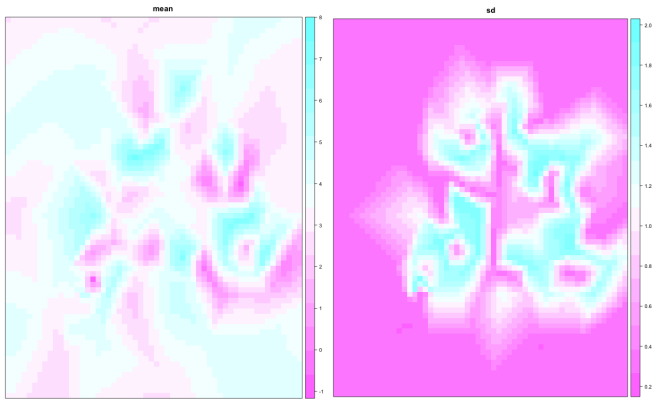


Los mapas se pueden obtener con el siguiente código:

```
print(levelplot(x=grid_mean,
row.values=proj_grid$x,
column.values=proj_grid$y,
col.regions=tim.colors(64),
   ylim=c(miny,maxy), xlim=c(minx,maxx),
aspect="iso",
contour=TRUE, cuts=21, labels=FALSE, pretty=TRUE,
xlab="Easting",ylab="Northing",
main=as.character(which_date)))
```

También se ha obtenido mapa para la media y desviación estándar del modelo. Podemos observar que la media varía entre el valor 3 y 5, aplicando exponencial sería entre 20 y 148 lo que es coherente con los valores de datos, por otro lado podemos ver que la desviación estándar en general es muy pequeña.

Figura 5.3.6: Media y desviación estándar



5.3.8. Conclusiones

La característica más sobresaliente de INLA es lo bajos tiempos computaciones que utiliza, en este caso no implica más de 2 horas con un computador de 2GB de ram y un procesador intel estándar. Actualmente se encuentran diversas ayuda de [R-INLA] la que se puede revisar para conseguir los resultados mostrados en el trabajo, en el caso particular de Santiago de Chile, la red de estaciones de monitoreo tiene solamente 11 estaciones, lo que proporcionar un mapa de estimación tiene sentido práctico.

Una linea de trabajo potencial es incorporar modelos meteorológicos que puedan modelar de mejor forma las covariables ambientales (temperatura, humedad relativa y velocidad del viento), así también resulta un futuro trabajo estudiar otras funciones base.

Capítulo 6

Apéndice.

6.1. Código de Tesis

6.1.1. Análisis construcción de base de datos

```
#PROMEDIO DIARIO
rm(list = ls())
# BASE DE DATOS DEF CONTIENE TODOS LAS COVARIABLES Y PM
Datos1 <- read.delim("base_def_NaN.txt",header=T, dec=",")
DAT<-as.matrix(Datos1)
### PROMEDIO DIARIO
PROM=c() for (a in 2012:2014){
for (m in 1:12){
for (d in 1:31){
ind=which(DAT[,4]==d & DAT[,3]==m & DAT[,2]==a)
PROM<-rbind(PROM,colMeans(DAT[ind,],na.rm=T))
} } }

## AHORA QUITAR DIAS 31 QUE NO EXISTEN
indNaN<-which(PROM[,4]=="NaN")
PROM.D<-PROM[-indNaN,]
S<-dim(PROM.D)
PROMEDIO12<-PROM.D
head(PROMEDIO12)
#CONSTRUCCIÓN DE MATRIZ
DATM<-c()
for (j in 1:S[1]){
for (i in 1:11){
DATM<-rbind(DATM,c(PROMEDIO12[j,2],
PROMEDIO12[j,3],
PROMEDIO12[j,4],
```

```
PROMEDIO12[j,(i*4+2)],
PROMEDIO12[j,(i*4+3)],
PROMEDIO12[j,(i*4+4)],
PROMEDIO12[j,(i*4+5)])) } }
DATAMATRIZ<-data.frame(DATM)
dim(DATAMATRIZ)
colnames(DATAMATRIZ) <- c("ANO", "MES" ,"DIA", "PM", "TEM", "HR", "VV")

# MATRIZ DEFINITIVA
MATRIZ_DEF<-cbind(COV,PM2.5=DATAMATRIZ[,4])

# GUARDAMOS BASE DE DATOS DEFINITIVA, ARREGLAR FECHAS Y ID POR ESTACION
save(MATRIZ_DEF, file="MATRIZ_DEF.RData")
head(MATRIZ_DEF) dim(MATRIZ_DEF) # DATAMATRIZ[c(1:60),]
fecha <- read.delim("fecha.txt",header=T)
ID<-rep(1:11,1095)
esta <- read.delim("estaciones_0.txt",header=T,dec=",")
COORDENADAS<-cbind(rep(esta[,2],1095),rep(esta[,3],1095))

#MATRIZ_FIN ES LA FINAL, CON FECHAS, ID POR ESTACION DE 3 AÑOS
MATRIZ_FIN<-cbind(Station.ID=ID,fecha,UTMX=COORDENADAS[,1],UTMY=COORDENADAS[,2],MATRIZ_DEF[,-c(1,2,3)])
head(MATRIZ_FIN)
MATRIZ_FIN[1:60,]
dim(as.matrix(MATRIZ_FIN))
save(MATRIZ_FIN, file = "MATRIZ_FIN.RData")

# MATRIZ DEFINITIVA PARA ALGORITMO (MATRIZ_DATOS)
ind3=which(MATRIZ_FIN[,1]==3 )
ind6=which(MATRIZ_FIN[,1]==6 )
ind8=which(MATRIZ_FIN[,1]==8)
MATRIZ_DATOS<-MATRIZ_FIN[-c(ind3,ind6,ind8),]

# DATOS_MATRIZ CONTIENE INFO NECESARIO PARA ALGORITMO
DATOS_MATRIZ<-data.frame(MATRIZ_DATOS)
head(DATOS_MATRIZ)
save(DATOS_MATRIZ, file = "DATOS_MATRIZ.RData")
write.table(DATOS_MATRIZ, file = "DATOS_MATRIZ.txt",sep="\t")

# MATRIZ DE VALIDACIÓN (MATRIZ_VAL)
ind1=which(MATRIZ_FIN[,1]==1 )
ind2=which(MATRIZ_FIN[,1]==2 )
ind4=which(MATRIZ_FIN[,1]==4)
ind5=which(MATRIZ_FIN[,1]==5)
ind7=which(MATRIZ_FIN[,1]==7)
ind9=which(MATRIZ_FIN[,1]==9)
ind10=which(MATRIZ_FIN[,1]==10)
ind11=which(MATRIZ_FIN[,1]==11)
MATRIZ_VAL<-MATRIZ_FIN[-c(ind1,ind2,ind4,ind5,ind7,ind9,ind10,ind11),]
```

```
save(MATRIZ_VAL, file = "MATRIZ_VAL.RData")
```

6.1.2. Modelo en R

```
# Cargamos paquetes necesarios
library(abind)
library(sp)
library(grid)
library(maps)
library(spam)
library(Matrix)
library(splines)
library(INLA)
library(fields)

# INICIO DEL ALGORITMO

#SELECCIONAR EL DIA A PREDECEDIR
if (!exists("i_dia")) { i_DIA = 200}
## LLAMADO A PROGRAMA SELECTOR DE COVARIABLES
if (!exists("pm10.path")) { pm10.path = "Covariates/" }

#CARGA DE LOS DATOS (8 ESTACIONES)
load("DATOS_MATRIZ.RData")
DATOS<-DATOS_MATRIZ
summary(DATOS)
head(DATOS)
DATOS[1:60,]

# COORDENADAS DE LAS 8 ESTACIONES
COORD_ESTACIONES<-read.delim("estaciones.txt",h=T,dec=",")

#DATOS DE VALIDACIÓN
load("MATRIZ_VAL.RData")
DATOS_VALIDACION<-MATRIZ_VAL

# COORDENADAS DE LAS ESTACIONES DE VALIDACION
COORD_VALIDACION<-read.delim("estaciones_val.txt",h=T,sep="\t",dec=",")
rownames(COORD_ESTACIONES)=COORD_ESTACIONES[,"Station.ID"]
rownames(COORD_VALIDACION) = COORD_VALIDACION[,"Station.ID"]

# DEFINIMOS MENSAJE EN MAPA
which_date = unique(DATOS$Date)[i_DIA]
print(paste("**---- T? puede predecir el d?a", which_date, "----**"))
# BORDE DE SANTIAGO DE CHILE
BORDE<-read.delim("Stgo.txt",h=T,sep="\t")

# CALCULOS INICIALES
```

```
N_ESTACIONES <- length(COORD_ESTACIONES$Station.ID)
N_ESTACIONES_VAL <- length(COORD_VALIDACION$Station.ID)
N_DATOS <- length(DATOS$Station.ID)
N_DIAS <- as.integer(N_DATOS/N_ESTACIONES)

# ESTANDARIZACION Y OTROS CALCULOS E INCOPORACIÓN DE LOG (PM_2.5) Y VECTOR DE TIEMPOS #
MEDIA_COVARIABLES=apply(DATOS[,3:7],2,mean,na.rm=T)
SD_COVARIABLES = apply(DATOS[,3:7],2,sd,na.rm=T)
DATOS[,3:7]=scale(DATOS[,3:7],MEDIA_COVARIABLES, SD_COVARIABLES)
DATOS_VALIDACION[,3:7]= scale(DATOS_VALIDACION[,3:7],MEDIA_COVARIABLES, SD_COVARIABLES)
DATOS$logPM2.5 = log(DATOS$PM2.5)
DATOS_VALIDACION$logPM2.5 = log(DATOS_VALIDACION$PM2.5)
DATOS$time = rep(1:N_DIAS,each = N_ESTACIONES)
DATOS_VALIDACION$time = rep(1:N_DIAS,each = N_ESTACIONES_VAL)

# TRIANGULACION DE LA REGION DE ESTUDIO
MESH = inla.mesh.create.helper(points=cbind(COORD_ESTACIONES$UTMX,COORD_ESTACIONES$UTMY),
points.domain=BORDE,n=5,offset=c(0.05,0.1), max.edge=c(0.3, 0.5), min.angle=c(26, 21))
par=mfrow=c(1,1)
plot(MESH)
lines(BORDE,type="p",cex=0.1)

#AGREGAMOS ESTACIONES
points(COORD_ESTACIONES$UTMX,COORD_ESTACIONES$UTMY,pch=20,cex=2,col=2)
points(COORD_VALIDACION$UTMX,COORD_VALIDACION$UTMY,pch=20,cex=2,col=4)

#CONSTRUCCION DE SPDE
spde = inla.spde2.matern(mesh=MESH, alpha=2)

#ESTRUCTURA PARA LOS DATOS
 A.est=inla.spde.make.A(MESH,loc=as.matrix(COORD_ESTACIONES[DATOS$Station.ID,
c("UTMX","UTMY")]),
group=DATOS$time,
n.group=N_DIAS
)

# ESTRUCTURA PARA DATOS DE VALIDACION
 A.val=inla.spde.make.A(MESH,loc=as.matrix(COORD_VALIDACION[DATOS_VALIDACION$Station.ID,
c("UTMX","UTMY")]),
group=DATOS_VALIDACION$time,
n.group=N_DIAS
)

# ESTRUCTURA PARA LA PREDICCIÓN
 A.pred=inla.spde.make.A(MESH, group=i_DIA, n.group=N_DIAS)
field.indices =inla.spde.make.index("field",
n.spde=spde$n.spde,
n.group=N_DIAS)
```

```
stack.est = inla.stack(data=list(logPM2.5=DATOS$logPM2.5),
A=list(A.est, 1),
effects=
list(c(field.indices,
list(Intercept=1)),
list(DATOS[,3:7])),
tag="est")

stack.val = inla.stack(data=list(logPM2.5=NA),A=list(A.val, 1),
                       effects=
                       list(c(field.indices,
                       list(Intercept=1)),
                       list(DATOS_VALIDACION[,3:7])),
                       tag="val")


scaled.mesh.loc = list(UTMX=(rep(scale(MESH$loc[,1],
                                       MEDIA_COVARIABLES["UTMX"],
                                       SD_COVARIABLES["UTMX"]),
                                       N_DIAS)),
                                       UTMY=(rep(scale(MESH$loc[,2],
                                       MEDIA_COVARIABLES["UTMY"],
                                       SD_COVARIABLES["UTMY"]),
                                        N_DIAS)))

stack.pred=inla.stack(data=list(logPM2.5=NA),
A=list(A.pred),  effects=list(c(field.indices,
                                       scaled.mesh.loc,
                                       list(Intercept=1) )),
                                       tag="pred")
stack = inla.stack(stack.est, stack.val, stack.pred)


# DEFINIMOS MODELO
    formula <- (logPM2.5 ~ -1 + Intercept + UTMX + UTMY + TEM + HR + VV +
                                f(field, model=spde, group=field.group, control.group=list(model="ar1")))

# EXTRAEMOS RESULTADOS DEL MODELO
if (!exists("do.remote")) { do.remote = FALSE }
    if (!exists("do.remote") || do.remote) {
 result =
inla(formula,
data=inla.stack.data(stack, spde=spde),
family="gaussian",
control.predictor=list(A=inla.stack.A(stack), compute=TRUE),
control.compute=list(cpo=FALSE),
control.inla = list(reordering = "metis"),
keep=FALSE, verbose=TRUE,
```

```
inla.call="remote", num.threads=12)
 }

else {
result =
inla(formula,
data=inla.stack.data(stack, spde=spde),
family="gaussian",
control.predictor=list(A=inla.stack.A(stack), compute=TRUE),
control.compute=list(cpo=FALSE),
control.inla = list(reordering = "metis"),
keep=FALSE, verbose=TRUE)     }
print(summary(result))

#GRÁFICA DE LA VARIANZA DEL MODELO
round(result$summary.fixed,4)
post.s2e<-inla.tmarginal(function (x) 1/x,
result$marginals.hyperpar$`Precision for the Gaussian observations`)
plot(post.s2e, type='l', ylab='Density',xlab=expression(sigma[e]^2))

#VECTOR DE PARÁMETROS
beta = result$summary.fixed[,"mean"]
beta  beta_sd = result$summary.fixed[,"sd"]

#INFORMACIÓN DE LOS DATOS DE VALIDACIÓN
validation0=
        list(p=rep(NA, length(DATOS$logPM2.5)))
index = inla.stack.index(stack,"est")$data
tmp.mean = result$summary.linear.predictor[index,"mean"]
tmp.sd = result$summary.linear.predictor[index,"sd"]
validation0$res = DATOS$logPM2.5 - tmp.mean
validation0$res.std = validation0$res /
sqrt(tmp.sd^2 + 1/result$summary.hyperpar[1,"mean"])
validation0$p = pnorm(validation0$res.std)

validation = list()
index = inla.stack.index(stack,"val")$data
tmp.mean = result$summary.linear.predictor[index,"mean"]
tmp.sd = result$summary.linear.predictor[index,"sd"]
validation$res = DATOS_VALIDACION$logPM2.5 - tmp.mean
validation$res.std = validation$res /
sqrt(tmp.sd^2 + 1/result$summary.hyperpar[1,"mean"])
validation$p = pnorm(validation$res.std)
validation$rmse = sqrt(mean(validation$res^2, na.rm=TRUE))
validation$cor = cor(DATOS_VALIDACION$logPM2.5, tmp.mean,
use="pairwise.complete.obs", method="pearson")
validation$cover=mean((validation$p >0.025)&(validation$p <0.975),na.rm=TRUE)
```

```
#RESULTADO DE SPDE
result.field = inla.spde.result(result, "field", spde, do.transform=TRUE)
field_mean = matrix(result.field$summary.values$mean, MESH$n, N_DIAS)
field_sd   = matrix(result.field$summary.values$sd, MESH$n, N_DIAS)
field_pred_mean = result$summary.linear.predictor[inla.stack.index(stack,
"pred")$data, "mean"]
field_pred_sd =     result$summary.linear.predictor[inla.stack.index(stack,
"pred")$data, "sd"]

# AR(1) PARÁMETROS
result$summary.hyperpar["GroupRho for field",]

# MATRIZ DE PRECISIÓN
sigma2eps_marg=inla.tmarginal(function(x) 1/x,
result$marginals.hyperpar$"Precision for the Gaussian observations")
sigma2eps_m1 = inla.emarginal(function(x) x, sigma2eps_marg)
sigma2eps_m2 = inla.emarginal(function(x) x^2, sigma2eps_marg)
sigma2eps_stdev = sqrt(sigma2eps_m2 - sigma2eps_m1^2)
sigma2eps_quantiles = inla.qmarginal(c(0.025, 0.5, 0.975), sigma2eps_marg)
var.nom.marg = result.field$marginals.variance.nominal[[1]]
var.nom.m1 = inla.emarginal(function(x) x,
var.nom.marg) var.nom.m2 = inla.emarginal(function(x) x^2,
var.nom.marg) var.nom.stdev = sqrt(var.nom.m2 - var.nom.m1^2)
var.nom.quantiles = inla.qmarginal(c(0.025, 0.5, 0.975), var.nom.marg)
range.nom.marg = result.field$marginals.range.nominal[[1]]
range.nom.m1 = inla.emarginal(function(x) x, range.nom.marg)
range.nom.m2 = inla.emarginal(function(x) x^2, range.nom.marg)
range.nom.stdev = sqrt(range.nom.m2 - range.nom.m1^2)
range.nom.quantiles = inla.qmarginal(c(0.025, 0.5, 0.975), range.nom.marg)
approx.hyperpar =     rbind(obs.var=c(sigma2eps_m1,
sigma2eps_stdev, sigma2eps_quantiles),
spde.var.nom=c(var.nom.m1, var.nom.stdev, var.nom.quantiles),
spde.range.nom=c(range.nom.m1, range.nom.stdev, range.nom.quantiles),
AR.rho=result$summary.hyperpar["GroupRho for field",1:5]) print(approx.hyperpar)

# GRILLAS DE CADA COVARIATE - CON MODELO LINEAL SIMPLE
load(paste(pm10.path,"DATOS_GRILLA.RData",sep=""))
load("COV_TEM.RData")
load("COV_HR.RData")
load("COV_VV.RData")

mean_covariates=MEDIA_COVARIABLES
sd_covariates=SD_COVARIABLES
covariate_array_std =    covariates_selector_funct(i_day=180,
MEDIA_COVARIABLES, SD_COVARIABLES)

#MAPA SOBRE GRILLA
minx<-min(COORD_ESTACIONES[,2])-0.3
```

```
maxx<-max(COORD_ESTACIONES[,2])+0.3
miny<-min(COORD_ESTACIONES[,3])-0.3
maxy<-max(COORD_ESTACIONES[,3])+0.3
grid1<-expand.grid(seq(minx,maxx,l=168),seq(miny,maxy,l=72))
plot(grid1)


proj_grid =      inla.mesh.projector(MESH,
xlim=range(grid1[,1]),
ylim=range(grid1[,2]),
    dims=c(56,72))

# LATENTE + COVARIABLES GEOGRÁFICAS
grid_latent_mean = inla.mesh.project(proj_grid, field_pred_mean)
grid_latent_sd = inla.mesh.project(proj_grid, field_pred_sd)

# MAPA DE PREDICCIÓN
trellis.par.set(regions=list(col=terrain.colors(16)))
grid.arrange(levelplot(grid_latent_mean ,
scales=list(draw=F), xlab='', ylab='', main='mean'),
levelplot(  grid_var ^(1/2), scal=list(draw=F), xla='', yla='', main='sd'), nrow=1)

library("lattice")
library("gridExtra")
print(levelplot(x=grid_mean,
row.values=proj_grid$x,
   column.values=proj_grid$y,
   col.regions=tim.colors(64),
  ylim=c(miny,maxy), xlim=c(minx,maxx),
          aspect="iso",
 contour=TRUE, cuts=21,
labels=FALSE, pretty=TRUE,
    xlab="Easting",ylab="Northing",
    main=as.character(which_date)))
trellis.focus("panel", 1, 1,
highlight=FALSE) lpoints(BORDE,col=1,cex=.25)
lpoints(COORD_ESTACIONES$UTMX, COORD_ESTACIONES$UTMY,col=1,lwd=2,pch=21)
```

6.2. Método de Laplace

Es un método que permite obtener distribuciones marginales a posteriori, propuesto por [Tierney and Kadane (1986)]. Este método aproxima numéricamente integrales de la siguiente forma:

$$\int_x exp\left\{f(x)\right\} dx \approx \sqrt{\frac{2\pi}{\mid f''(x_0) \mid}} exp\left\{f(x_0)\right\}$$

donde x_0 es el punto máximo de la función $f(x)$ y $f''(x_0)$ es la segunda derivada de la función $f(x)$ evaluada en x_0.

6.2.1. Ejemplo binomial

Sea Y una variable aleatoria con distribución binomial ($x \sim Bin(n,p)$) con $n$fijo y probabilidad p.

Sea $\eta = ln\left(\frac{p}{1-p}\right)$, entonces podemos reescribir p por $\left(\frac{exp(\eta)}{1+exp(\eta)}\right)$. La función de verosimilitud de k observaciones está dada por:

$$L(y;\eta) = \prod_{i=1}^{k} C_{y_i}^{n} \frac{exp(\eta)^{y_i}}{(1+exp(\eta))^{y_i}} \left(1 - \frac{exp(\eta)}{1+exp(\eta)}\right)^{(n-y_i)}$$

$$= \left(\prod_{i=1}^{k} C_{y_i}^{n}\right) exp(\eta) \sum_{i=1}^{k} y_i \left(1+exp(\eta)\right)^{-nk}$$

Considerando que $0 \leq p \leq 1$, usando que η se distribuye normal (μ, σ^2) cuya pdf está está dada por:

$$\pi(\eta) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2\sigma^2}\left(\eta-\mu\right)^2\right\}$$

Usando el teorema de Bayes, se puede encontrar la distribución a posteriori conjunta para η:

$$\pi\left(\eta \mid y\right) = \frac{\pi\left(\eta\right) L(y;\eta)}{\int_{-\infty}^{\infty} \pi(\eta) L(y;n) d\eta}$$

Sustituyendo y simplificando se obtiene:

$$\pi(\eta \mid y) = \frac{exp(\eta)^{\sum_{i=1}^{k} y_i} \left(1+exp(\eta)\right)^{-nk} exp\left\{-\frac{1}{2\sigma^2}\left(\eta-\mu\right)^2\right\}}{\int_{-\infty}^{\infty} exp(\eta)^{\sum_{i=1}^{k} y_i} \left(1+exp(\eta)\right)^{-nk} exp\left\{-\frac{1}{2\sigma^2}(\eta-\mu)^2\right\} d\eta}$$

Es necesario resolver la integral del denominador:

$$\int_{-\infty}^{\infty} exp(\eta)^{\sum_{i=1}^{k}} (1+exp(\eta))^{-nk} exp\left\{-\frac{1}{2\sigma^2}(\eta-\mu)^2\right\} d\eta$$

Esta integral no puede resolverse analíticamente, entonces aplicamos método de Laplace.

$$f(\eta) = log\left(exp(\eta)^{\sum_{i=1}^{k}} (1+exp(\eta))^{-nk} exp\left\{-\frac{1}{2\sigma^2}(\eta-\mu)^2\right\}\right)$$

Tal como describe el método aplicamos aproximación numérica en punto η_0.

$$f^{'}(\eta) = \sum_{i=1}^{k} y_i - nk\frac{exp(\eta)}{(1+exp(\eta))} - \frac{1}{\sigma^2}(\eta-\mu)$$

Analíticamente no es posible encontrar el punto máximo de esta función, pero supondremos que el punto máximo η_0 puede ser encontrado mediante un método por ejemplo newton Rapson.

Calculamos la segunda derivada de $f(\eta)$

$$f^{''}(\eta) = -nk\frac{exp(\eta)}{(1+exp(\eta))^2} - \frac{1}{\sigma^2}$$

sea

$$\gamma_0 = \frac{1}{\sqrt{\mid f^{''}(x) \mid}}$$

$$\gamma_0 = \frac{1}{\sqrt{\left|\left(-nk\frac{exp(\eta_0)}{(1+exp(\eta_0))^2} - \frac{1}{\sigma^2}\right)\right|}}$$

Así tenemos que:

$$\int_{-\infty}^{\infty} exp(\eta)^{\sum_{i=1}^{k} y_i} (1+exp(\eta))^{-nk} exp\left\{-\frac{1}{2\sigma^2}(\eta-\mu)^2\right\} d\eta$$

$$\approx \sqrt{2\pi}\gamma_0 exp\left\{\eta_0 \sum_{i=1}^{k} y_i - nk\, ln(1+exp(\eta_0)) - \frac{1}{2\sigma^2}(\eta_0-\mu)^2\right\}$$

Resolviendo tenemos que:

$$\pi(\eta \mid y) \approx \frac{exp\left\{\eta \sum_{i=1}^{k} y_i - nk\, ln\left(1 + exp(\eta)\right) - \frac{1}{2\sigma^2}(\eta - \mu)^2\right\}}{\sqrt{2\pi}\gamma_0 exp\left\{\eta_0 \sum_{i=1}^{k} y_i - nk\, ln\left(1 + exp(\eta_0)\right) - \frac{1}{2\sigma^2}\left(\eta_0 - \mu\right)^2\right\}}$$

Entonces la distribución de p puede ser obtenida mediante jacobiano, ya que $p = \frac{exp(\eta_0)}{(1 + exp(\eta))}$

$$\pi\left(p \mid y\right) \approx \frac{exp\left\{ln\left(\frac{p}{1-p}\right)\sum_{i=1}^{k} y_i - nk\left(1 + \frac{p}{1-p}\right) - \frac{1}{2\sigma^2}\left(ln\left(\frac{p}{1-p}\right) - \mu\right)^2\right\}}{p\left(1-p\right)\sqrt{2\pi}\gamma_0 exp\left\{\eta_0 \sum_{i=1}^{k} y_i - nk\, ln\left(1 + exp(\eta_0)\right) - \frac{1}{2\sigma^2}\left(\eta_0 - \mu\right)^2\right\}}$$

6.3. Ejemplos de R -inla

6.3.1. Modelo binominal

Sea $Y_1, Y_2, ..., Y_k$ una muestra aleatoria, tal que $y_i \sim Bin(n, p)$ donde el parámetro a descubrir la distribución es p.

bajo el enfoque bayesiano podemos asumir una distribución apriori para p.

Aquí tenemos un solo parámetro $x = \{p\}$, precisamos una distribución normal para x. Como tenemos que $(p \in [0, 1])$, consideraremos lo siguiente

$$\eta = log\left(\frac{p}{1-p}\right) = x$$

ahora podemos reescribir el parámetro de la siguiente forma $\frac{exp(\eta)}{(1 + exp(\eta))}$

Esta transformación también permite trabajar adecuadamente con un modelo aditivo generalizado .

En R ya tenemos esto implementado y haremos uso de ello.

Fijaremos una semilla aleatoria para que los primeros resultados se puedan repetir y generará una muestra de tamaño 100, $n = 1$, $p = 0{,}2$. El parámetro n puede interpretarse como número ensayos, $p$puede ser interpretado como la probabilidad de tener éxito en cada ensayo.

Aquí se considera $\tau = \frac{1}{\sigma^2}$, $\eta \sim N\left(0, \frac{1}{\sigma^2}\right)$. (parametrizaciones ya que R usa a priori con media y precisión τ para modelo de efecto fijo).

```
# Ejemplo de binomial
rm(list = ls()); p=0.2; n=1; k=100
# semilla
set.seed(4+8+15+16+23+42)
y=rbinom(k,n,p)
data=data.frame(y)
sigma2=1000
#Precisión
presicion=1/sigma2
#
formula=y~1
resultado=inla(formula,data=data,family = "binomial",
Ntrials = n,control.fixed = list(prec.intercept=presicion))
```

6.3.2. SPDE. Triangulación

La estructura general es:

```
inla.mesh.2d(loc = NULL,
loc.domain = NULL,
offset = NULL,
n = NULL,
boundary = NULL,
interior = NULL,
max.edge,
min.angle = NULL,
cutoff = 1e-12,
plot.delay = NULL)
```

Se necesita una región, que se puede especificar la ubicación, puntos o dominio. La localización específica en loc, se utilizan los nodos iniciales de triangulación.

Un polígono puede ser definido por un dominio en $loc.domain$. Si proporcionamos las localizaciones de los puntos o bien entregamos un dominio con el argumento $loc.domain$ el algoritmo y una malla convexa. Una malla no convexa se puede hacer cuando proporcionamos un conjunto de límites de un polígono $inla.mesh.segment$.

Una de las clases detalladas es obligatoria. El otro argumento obligatorio es $max.edge()$, este argumento especifica la máxima longitud de las aristas del interior y exterior.

Los otros argumentos permiten especificar condiciones adicionales. El $offset$ es numérico o vector. Si es negativo es interpretado como un factor en relación con el diámetro aproximado

de datos (por defecto = -0.10. Si es positivo representa la distancia interior y exterior de la extensión.

El argumento $n$es el número inicial de puntos del contorno, el interior es una lista para especificar el contorno cada uno de los $inla.mesh$.

Una buena malla de triángulos tiene que tener lo más regulares posibles en tamaño y forma. Para lograr esto se puede manipular tanto $max.edge$ tenemos el argumento $min.angle$ que puede ser escalar o vectorial, y permite especificar los ángulos interiores de los triángulos del interior del borde y exterior. Los valores hasta el 21 garantizan la convergencia del algoritmo.

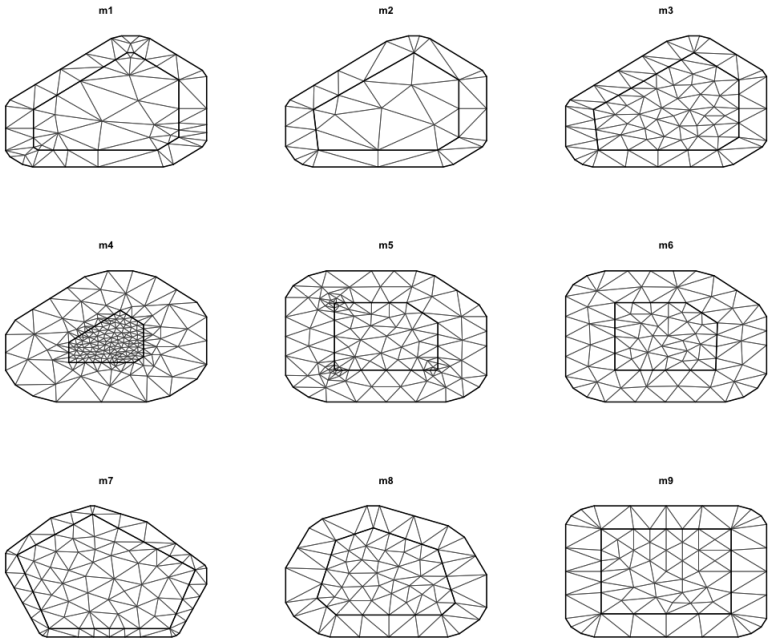
Adicionalmente se puede usar el argumento de $cutoff$, lo cual es la distancia mínima permitida entre puntos. Esto significa que los puntos a una distancia más cercana que el valor definido se sustituye por un único vértice. De esta manera se evitan pequeños triángulos y debe ser un número positivo y es crítico cuando tenemos muy cercanos.

6.3.2.1. Ejemplo SPDE

```
library(INLA);library(fields)
data(SPDEtoy)
head(SPDEtoy)
coords <- as.matrix(SPDEtoy[,1:2]) ; p5 <- coords[1:5,]
pl.dom <- cbind(c(0,1,1,0.7,0), c(0,0,0.7,1,1))
m1 <- inla.mesh.2d(p5, max.edge=c(0.5, 0.5))
m2 <- inla.mesh.2d(p5, max.edge=c(0.5, 0.5), cutoff=0.1)
m3 <- inla.mesh.2d(p5, max.edge=c(0.1, 0.5), cutoff=0.1)
m4 <- inla.mesh.2d(p5, max.edge=c(0.1, 0.5), offset=c(0,-0.65))
m5 <- inla.mesh.2d(, pl.dom, max.edge=c(0.3, 0.5), offset=c(0.03, 0.5))
m6 <- inla.mesh.2d(, pl.dom, max.edge=c(0.3, 0.5), offset=c(0.03, 0.5), cutoff=0.1)
m7 <- inla.mesh.2d(, pl.dom, max.edge=c(0.3, 0.5), n=5, offset=c(.05,.1))
m8 <- inla.mesh.2d(, pl.dom, max.edge=c(.3, 0.5), n=7, offset=c(.01,.3))
m9 <- inla.mesh.2d(, pl.dom, max.edge=c(.3, 0.5), n=4, offset=c(.05,.3))
# n=j, número vértices del polígono.
par(mfrow=c(3,3))
plot(m1); plot(m2); plot(m3)
plot(m4); plot(m4); plot(m6)
plot(m7); plot(m8); plot(m9)
```

La gráfica obtenida es la siguiente:

Figura 6.3.1: Gráficas de triangulación.



En $m1$ hay dos problemas, algunos triángulos con ángulos pequeños (al interior) y algunos triángulos en el dominio interior hay triángulos grandes.

En $m2$ se ha suavizado la restricciones en las locaciones por que los puntos a una menor distancia se consideran como un solo punto, esto evita que algunos de los triangular(lado superior derecho) con ángulos pequeños como $m1$.

Cada uno de los triángulos interiores de $m3$ en la parte superior derecha tiene un longitud menor a 0.1 entonces esta malla luce mucho mejor que los dos anteriores. El $m4$ fue construido sin primero construir la malla convexa de los puntos de la región. Sólo tenia el segundo contorno. En este caso longitud de los triángulos no funciona (el primer valor en el argumento $max.edge$) . En $m5$ solamente se define usando el dominio del polígono. En esta malla tenemos algunos

triángulos pequeños en las esquinas, debido a que fue construido sin especificar el límite. Tenemos extrañas regiones trianguladas muy pequeñas. En $m6$ se ha añadido el limite y estuvieron una malla mucho mejor que la anterior.

En las últimas 3 mallas cambiaron el número inicial de los puntos de extensión esto puede ser útil en lagunas situaciones para obtener convergencia. Aquí muestran la forma de la malla que obtuvieron por ejemplo n=5 en $m7$. Este número produce una malla que parece inadecuada para este dominio por que tienen una extensión no uniforme detrás del borde. Finalmente $m9$ la forma de los triángulos es muy mala.

Bibliografía

[Banerjee et al (2004)] Banerjee, S., Carlin, B.P. y Gelfand, A.E. (2004). Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC. Monographs on Statistics & Applied Probability.

[CONAMA (2010)] Comisión Nacional Del Medio Ambiente (CONAMA), 2010. Informe final relación de la norma de calidad primaria MP2,5 con la norma de calidad primaria de MP10. Preparado por Luis Cifuentes. Santiago: Conama.

[Cressie (1993)] Cressie N (1993) Statistics for Spatial Data. Wiley, New York

[Cameletti et al (2012)] Cameletti, M., Lindgren, F., Simpson, D. and Rue, H., 2012. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. AStA Advances in Statistical Analysis, 97 (2), pp. 109-131

[Lindgren et al (2011)] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach (with discussion). Journal of the Royal Statistical Society. Series B. Statistical Methodology, 73(4):423–498, September 2011.

[MMA (2011)] Ministerio del Medio Ambiente (MMA), 2011a. Análisis general de impacto económico y social del anteproyecto de revisión de la

norma de emi- sión de NO, HC y CO para el control del NOx en vehículos en uso, de encendido por chispa (AGIES). Santiago: MMA.

[Matérn (1986)] Matérn, B. (1986). Spatial Variation. Springer-Verlag Berlin and Heidelberg GmbH & Co.

[Rue y Held (2005)] Rue, H. y Held, L. (2005). Gaussian Markov Random Fields; Theory and Applications, vol. 104 of Monographs on Statistics and Applied Probability, Chapman & Hall/CRC.

[Rue et al (2009)] Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian model by using integrated nested Laplace approximations (with discussion). J R Statist Soc B 71:319–392

[R-INLA] http://www. r-inla.org.

[R team (2010)] R Development Core Team (2010). R: A language and environment for statistical comput- ing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[Stein (1999)] Stein, M.L. (1999). Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

[Tierney and Kadane (1986)] Tierney L and Kadane J (1986). Accurate Aproximations for Posterior Moments and Marginal Densities, Journal of the American Statistical Associations, 81, 82-86

[Vargas (2011)] Vargas C. (2011). Efectos de la fracción gruesa (PM10-2.5) del material particulado sobre la salud humana. Revisión Bibliográfica. MINSAL.

[Whittle (1963)] Whittle, P. (1963). Stochastic process in several dimensions. Bulletin of the International Statistical Institute, 40, 974-985.

[Wanckernagel (1995)] Wanckernagel H, (1995) Multivariate Geostatistics. An Introduction with Applications. Springer-Verlag, Berlin.