Universidad ECCI
Departamento de Ciencias Básicas
Pregrado en Estadística

Modelo predictivo de difusión de contaminante $PM_{2,5}$ en el área urbana de Bogotá entre 2017-2020 en presencia del Covid-19

Tesis para optar al grado de Profesional en Estadística

Gerson José Moreno Morales
Bogotá-Colombia
2024

Director: Pitter Javier Cabezas
Codirector: José Alexander Fuentes
Departamento de Ciencias Básicas
Universidad ECCI, Bogotá

Modelo predictivo de difusión de contaminante $PM_{2,5}$ en el área urbana de Bogotá entre 2017-2020 en presencia del Covid-19

Gerson José Moreno Morales

Director de Tesis: Pitter Cabezas, Universidad ECCI, Bogotá-Colombia.
Codirector de Tesis: Alexander Fuentes, Universidad EAN, Bogotá-Colombia.
Director de Programa: Laura Marcela Rúa Yáñez.

Jurados de Tesis

.....

.....

.....

Calificación: _____

Bogotá-Colombia, Mayo de 2024

Abstract

As a contribution to the study of the quality of breathing air and the need to quantify its general and specific characteristics, this dissertation discusses the diffusion of polluting material from static and mobile sources recorded in the network of air quality stations of Bogotá, postulating a spatio-temporal predictive model based on imputation (MissForest) and prediction (Spatial Autoregressive Hilbert ($SARH(1)$)) methodologies. These procedures made it possible to spatially configure the domain of the diffusion variables, as well as generate the concentration surfaces of particulate matter ($PM_{2,5}$) and its temporal evolution, predicting the respective values of the missing stations in the city. From this approach, the prediction efficiency of the model and its applicability in financial, economic, biological, environmental phenomena, among others, with diffusive characteristics and low resolution, are concluded.

Keywords— MissForest, Spatial Autorregressive Hilbert, particulate matter, diffusion, Hilbert space

Resumen

Como contribución al estudio de la calidad del aire respirable y la necesidad de cuantificar sus características generales y específicas, en el presente trabajo de grado se discute la difusión de material contaminante proveniente de fuentes estáticas y móviles registrado en la red de estaciones de calidad del aire de Bogotá, postulando un modelo predictivo espacio-temporal a partir de las metodologías de imputación (MissForest) y de predicción (Spatial Autorregressive Hilbert ($SARH(1)$)). Estos procedimientos permitieron configurar espacialmente el dominio de las variables de difusión, así como, generar las superficies de concentración de material particulado ($PM_{2,5}$) y su evolución temporal, prediciendo los valores respectivos de las estaciones faltantes en la ciudad. A partir de esta aproximación se concluye la eficiencia de predicción del modelo y su aplicabilidad en fenómenos financieros, económicos, biológicos, ambientales, entre otros, con características difusivas y baja resolución.

Palabras Claves— MissForest, Spatial Autorregressive Hilbert, material particulado, difusión, espacio de Hilbert

Agradecimientos

No fue un camino fácil pero, estuvo acompañado de grandes personas que me apoyaron en este proceso. No queda más que agradecer principalmente a mis grandes amigos Alexander y Pitter que han sido los promotores de la construcción de mi conocimiento hoy en día. También a mi esposa Luisa por su comprensión y paciencia. A Jaime y a Rosa que me dieron mucho aliento todos estos años. Finalmente, a cada una de las personas que con una conversación, una visita o una llamada me alentaron a alcanzar este logro.

Infinitas gracias a todos.

Índice general

Índice de cuadros

Índice de figuras

Introducción

En los campos del conocimiento existen diversas problemáticas que se relacionan implícitamente pues presentan comportamientos y características similares. Un caso particular de estas problemáticas relacionadas es en donde se involucran variables que se comportan como una sustancia difundiéndose a través de un fluido, como el caso de la concentración de material particulado en el aire del área metropolitana de Bogotá Instituto de Hidrología [2012]. La necesidad creciente de la predicción óptima de estas variables como soporte para la toma de decisiones en temas preventivos, correctivos, inversión, riesgo, entre otras, direcciona a la búsqueda de soluciones estadísticas. La incertidumbre juega un papel fundamental en estas soluciones, por tanto, el objetivo es minimizarla tanto como sea posible. La estadística, gracias a los aportes de algunos autores, ha adoptado diferentes modelos de predicción que cumplen con esta minimización de la incertidumbre. Dicho esto, para variables con las características mencionadas, se proponen, como modelos adecuados, los modelos espacio temporales y, ahora en la era digital, técnicas de Machine Learning.

En la rama del modelado estadístico Ruiz-Medina [2011] en su trabajo *Spatial functional prediction from spatial autoregressive Hilbertian processes*, propone un modelo de predicción espaciotemporal funcional basada en la diagonalización del parámetro de un proceso $SARH(1)$, el cual lo aplica en unos datos simulados y unos reales. En los datos reales se tiene la temperatura media anual de la superficie del océano en diferentes locaciones. En la validación cruzada encuentra que el Error Absoluto Medio se encuentra en el orden de 10-1 que es relativamente alto y no suficiente para tener confiabilidad en la predicción de esta variable. Luego, Ruiz-Medina and Espejo [2012], en su trabajo colaborativo *Spatial autoregressive functional plug-in prediction of ocean surface temperature*, juntan sus esfuerzos para continuar en la búsqueda de una confiabilidad más alta y proponen un complemento a su anterior modelo. En este calculan los estimadores basados en momentos de los operadores presentes en el proceso por medio de proyección en una base ortogonal adecuada considerando la diagonalización de la base de la función propia del operador de autocovarianza. Así, con datos recolectados por las estaciones meteorológicas ubicadas en el océano hawaiano, el modelo complementado refleja en la validación cruzada Errores Absolutos Medios en el orden de 10^{-2} en su mayoría de nodos, siendo así un modelo que minimiza aún más la incertidumbre, indicando que se puede mejorar aún más.

Por otro lado, Kaminska [2018] en su trabajo *The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław*, utiliza una técnica de machine learning llamada Random Forest para modelar la contaminación del aire dependiendo de las condiciones meteorológicas y el tráfico en Breslavia, Polonia. Sus resultados mejoraron al seleccionar las variables predictoras de su modelo de regresión por medio de esta técnica, ya que el coeficiente de correlación de Pearson no permitía identificar relaciones implícitas. Sin embargo, Mason et al. [1999] en su trabajo *Boosting Algorithms as Gradient Descent*, presentan un algoritmo de potenciación del gradiente funcional que optimizan bastante las predicciones o selecciones utilizando un conjunto de modelos de predicción débiles como lo es Random Forest.

Una de las preocupaciones más urgentes en la actualidad, en materia ambiental, corresponde a la determinación de los mecanismos de difusión de contaminantes que, al interactuar con la atmósfera, contribuyen a la calidad del aire respirable en zonas urbanas. En la ciudad de Bogotá, Colombia, existe una necesidad creciente acerca de la calidad del aire respirable que, en el caso de poseer agentes contaminantes, deteriora la salud y el

bienestar de la población capitalina (Secretaría Distrital de Ambiente de Bogotá [2021], Organización Mundial de la Salud [2021]). La captura de información a partir de la red de estaciones dispuestas para la medición del material particulado presente en el aire posee la dificultad de baja resolución espacial, debido a la poca cantidad instalada (20 en total), dispuestas no simétricamente a lo largo del área urbana. Adicionalmente, esta dificultad aumenta debido a las características y condiciones geográficas de la capital colombiana. Dicho esto, se hace necesario efectuar procedimientos predictivos de los mecanismos de la difusión de los contaminantes provenientes de fuentes móviles (relacionadas con automotores) y fuentes fijas (asociadas a las emisiones desde fábricas y procesos industriales), como soporte para la toma de decisiones frente a políticas ambientales, de salud, económicas, entre otras. Existe un indicador llamado ICA (Índice de Calidad del Aire) (Instituto de Hidrología [2012]), que por franjas de colores permite visualizar el rango adimensional en el que se encuentra la concentración de los diferentes contaminantes atmosféricos, los cuales influyen en la calidad del aire. Este indicador es creado por el IDEAM y es referente para que en el POT 2022 - 2030 "Bogotá Reverdece"(Secretaría Distrital de Ambiente de Bogotá [2021]) se modifique el IBOCA (Índice Bogotano de Calidad del Áire) en miras de abordar incidentes de manera preventiva y correctiva de ser el caso.

Se propone mediante un modelo predictivo funcional espacial autorregresivo hilbertiano de orden 1 ($SARH(1)$, Bosq [2012], Ruiz-Medina [2012], Ruiz and Angulo [2007]) y, por medio de la técnica no paramétrica de imputación de datos usando Random Forest (MissForest, Stekhoven and Bühlmann [2011]), efectuar un procedimiento iterativo que genere los datos de contaminación en zonas alejadas de las estaciones de medición, realizando suavizamiento vía B-spline, con el objetivo de obtener una estimación espacio temporal alternativa a los modelos individuales propuestos hasta el momento. Lo anterior, asociado a la disminución de las emisiones debido a los confinamientos efectuados a razón de la pandemia del Covid-19. En este caso específico, se toma, como contaminante, al $PM_{2,5}$ para predecir su difusión en el área metropolitana de Bogotá entre 2017 y 2020. Así mismo, la integralidad de este modelo permite que se pueda implementar en la predicción de cualquier variable que cumpla con las características espaciotemporales y de difusión, asociando variables que suplan las condiciones de la difusión.

Planteamiento

Planteamiento del problema

A través del tiempo la calidad del aire se ha catalogado como una de las principales amenazas de muerte a nivel mundial. Una exposición constante o de forma crónica a este tipo de $PM_{2,5}$ puede desarrollar enfermedades de tipo respiratorio o cardiovascular. La red de estaciones de monitoreo de la calidad del aire en Bogotá no está logrando proporcionar una representación precisa y completa de las concentraciones de $PM_{2,5}$ en toda la ciudad. La ubicación y cantidad de estaciones actuales pueden no ser suficientes para captar la variabilidad espacial de estas partículas finas, lo que limita la capacidad de los responsables de la toma de decisiones para implementar medidas específicas en áreas críticas.

Objetivos

Objetivo General

- Establecer un modelo predictivo para variables asociadas a problemáticas con comportamientos de difusión y generar un ejemplo para la ventana de 2020-2021 correspondiente a la presencia del confinamiento COVID-19 cuantificando el nivel de reducción de contaminante en el aire.

Objetivos Específicos

- Determinar la superficie de mayor predicción de difusión de contaminantes a partir de la combinación de las metodologías MissForest y $SARH(1)$ (Spatial autoregressive Hilbertian process).
- Predecir las características de la difusión de contaminantes en el área metropolitana de Bogotá a partir de los datos existentes en el período 2017-2020.

Metodología

Contextualización

En el ámbito medioambiental actual, una de las preocupaciones más apremiantes radica en la identificación de los mecanismos de difusión de contaminantes que, al interactuar con la atmósfera, inciden directamente en la calidad del aire respirable en entornos urbanos. La ciudad de Bogotá, Colombia, enfrenta una creciente inquietud respecto a la calidad del aire respirable, cuya contaminación potencialmente afecta la salud y el bienestar de la población capitalina.

El desafío principal reside en la recopilación de información mediante la red de estaciones diseñadas para medir el material particulado presente en el aire. Esta red, compuesta por un total de 16 estaciones, presenta limitaciones significativas, siendo su baja resolución espacial el principal obstáculo. La disposición no simétrica de estas estaciones a lo largo del área urbana de Bogotá dificulta la captura precisa de la variabilidad espacial de contaminantes. A esto se suma la influencia de las características geográficas específicas de la capital colombiana, que intensifican las dificultades en el monitoreo.

En respuesta a estas limitaciones, resulta imperativo desarrollar procedimientos predictivos que aborden los mecanismos de difusión de contaminantes, especialmente aquellos provenientes de fuentes móviles, como vehículos automotores, y fuentes fijas, asociadas a emisiones industriales. Con el objetivo de respaldar decisiones en políticas ambientales, de salud y económicas, se propone la implementación de un modelo predictivo funcional espacial autorregresivo hilbertiano de orden 1 ($SARH(1)$, Bosq [2012], Ruiz-Medina [2012], Ruiz and Angulo [2007]).

Para optimizar la precisión y alcance de las predicciones, se incorpora la técnica no paramétrica de imputación de datos mediante Random Forest (MissForest, Stekhoven and Bühlmann [2011]). Este enfoque se ejecuta en un procedimiento iterativo que genera datos de contaminación en áreas alejadas de las estaciones de monitoreo, con un suavizamiento adicional a través de P-spline. El propósito de este procedimiento es obtener una estimación espacio-temporal alternativa a los modelos individuales existentes hasta la fecha, considerando además la disminución de emisiones debido a las restricciones impuestas durante la pandemia del Covid-19.

En este caso específico, se selecciona el $PM_{2,5}$ como el contaminante de interés para predecir su difusión en el área metropolitana de Bogotá en el periodo comprendido entre 2017 y 2020. Vale la pena destacar que la versatilidad de este modelo posibilita su aplicación en la predicción de cualquier variable que cumpla con características espaciotemporales y de difusión, al asociar variables que satisfacen las condiciones inherentes al proceso de difusión en cuestión.

Descripción de la región de estudio

El área de Bogotá cubre 1.775 km^2 con un área urbana de 307 km^2. La topografía de la región se caracteriza por un terreno complejo limitada por la cordillera central al oriente. Está situado en un valle montañoso a 2.600 m sobre el nivel del mar.

Dataset

El conjunto de datos utilizado se obtuvo de la red de monitoreo de calidad del aire de Bogotá, cada hora, entre los años 2017-2020. La red de seguimiento consta de 20 estaciones como lo muestra la Figura 2 y la Tabla 1.

Estación	Latitud	Longitud	Altitud	Localidad	Dirección	Tipo de zona	Tipo de estación
Convenio 176	4° 45'2.15"N	74° 1'1.28"W	2575 m	Usaquén	Autopista Norte # 174-10	Urbana	De fondo
Usaquén	4° 42'37.26"N	74° 1'49.50"W	2570 m	Usaquén	Carrera 7B Bis # 132-11	Urbana	De fondo
Suba	4° 45'40.49"N	74° 5'36.46"W	2571 m	Suba	Carrera 111 # 159A-61	Suburbana	De fondo
Bolivia	4° 44'9.12"N	74° 7'33.18"W	2574 m	Engativá	Avenida Calle 80# 121-98	Suburbana	De fondo
Las Ferias	4° 41'26.52"N	74° 4'56.94"W	2552 m	Engativá	Avenida Calle 80# 69Q-50	Urbana	De tráfico
P. Simón Bolívar	4° 39'30.48"N	74° 5'2.28"W	2577 m	Barrios Unidos	Calle 63# 59A-06	Urbana	De fondo
Sagrado Corazón	4° 37'31.75"N	74° 4'1.13"W	2621 m	Santa Fe	Calle 37# 8-40	Urbana	De tráfico
Fontibón	4° 40'12.36"N	74°8'29.58"W	2591 m	Fontibón	Carrera 96G # 17B-49	Urbana	Industrial
Puente Aranda	4° 37'54.36"N	74° 7'2.94"W	2590 m	Puente Aranda	Calle 10# 65-28	Urbana	Industrial
Kennedy	4° 37'30.18"N	74° 9'40.80"W	2580 m	Kennedy	Carrera 80# 40-55 sur	Urbana	De fondo
Carvajal	4° 35'44.22"N	74° 8'54.90"W	2563 m	Kennedy	Autopista Sur # 63-40	Urbana	Industrial
Bosa	4° 65'14.2"N	74° 2'4.7"W	2563 m	La Libertad	Carrera 88c # 61a sur-14	Urbana	De fondo
Tunal	4° 34'34.41"N	74° 7'51.44"W	2589 m	Tunjuelito	Carrera 24# 49-86 sur	Urbana	De fondo, Industrial
San Cristóbal	4° 34'21.19"N	74° 5'1.73"W	2688 m	San Cristóbal	Carrera 2 Este# 12-78 sur	Urbana	De fondo
Usme	4° 35'23"N	74° 6.3"W	2581 m	Monte blanco	Carrera 14 # 96-8 sur	Urbana	De fondo

Figura 1: Ubicación geográfica de estaciones meteorológicas de la Sabana de Bogotá. Se incluye la localización, el tipo de zona y su categoría. Fuente: `http://ambientebogota.gov.co/estaciones-rmcab`

Figura 2: Ubicación geográfica de las estaciones meteorológicas de Bogotá.

Estas estaciones midieron, con los métodos estándar de *Thermo Fisher Scientific EPA*, tanto datos meteorológicos (temperatura, presión, velocidad del viento, dirección del viento, precipitaciones, radiación solar y humedad relativa) como contaminantes atmosféricos: monóxido de carbono (CO, $mg \cdot m^{-3}$), dióxido de azufre (SO_2, $\mu g \cdot m^{-3}$), ozono (O_3, $\mu g \cdot m^{-3}$), dióxido de nitrógeno (NO_2, $\mu g \cdot m^{-3}$), y material particulado $< 2{,}5$ μm ($PM_{2,5}$, $\mu g \cdot m^{-3}$).

Fundamentación Teórica

En diversos estudios de investigación sobre calidad del aire, se proponen modelos puntuales, lineales o que no describen la correlación espacial, tanto global, como a pequeña escala. Para proponer un modelo de contaminación del aire espacio-temporal, este enfoque abarca los pasos a continuación:

MissForest: Teoría, Algoritmo y Evaluación

Teoría

La técnica MissForest se basa en el uso de bosques aleatorios para imputar valores faltantes en conjuntos de datos. Utiliza un enfoque iterativo para adaptar el algoritmo de bosques aleatorios a datos con valores ausentes. A continuación, se describen los aspectos teóricos clave:

Bosques Aleatorios (Random Forest)

El algoritmo de bosques aleatorios se compone de múltiples árboles de decisión. Cada árbol se entrena con un conjunto de datos bootstrap y utiliza una selección aleatoria de variables en cada división de nodo. La predicción final se obtiene mediante la agregación de las predicciones individuales de cada árbol (Breiman [2001]).

MissForest: Proceso Iterativo

El proceso de MissForest se realiza iterativamente:

1. **Inicialización:** Se inicializan los valores faltantes utilizando algún método simple (por ejemplo, imputación por la media).
2. **Iteración del Bosque Aleatorio:** Para cada árbol en el bosque aleatorio:
   - Se selecciona aleatoriamente un subconjunto de variables que contienen valores faltantes.
   - El árbol se entrena para predecir los valores faltantes basándose en las variables observadas.
3. **Actualización de Imputaciones:** Se actualizan los valores faltantes utilizando las predicciones del bosque aleatorio.
4. **Criterio de Convergencia:** Se evalúa la diferencia entre las imputaciones actuales y las anteriores. El proceso se detiene cuando se alcanza la convergencia.

Algoritmo de MissForest

La imputación de MissForest para un conjunto de datos X con valores faltantes se realiza mediante el siguiente algoritmo:

1. **Inicialización:** Inicializar valores faltantes utilizando métodos simples.
2. **Para cada iteración:**
   - Seleccionar aleatoriamente un subconjunto de variables con valores faltantes.

- Entrenar un árbol de decisión para predecir los valores faltantes basándose en las variables observadas.
- Actualizar los valores faltantes con las predicciones del árbol.

3. **Detenerse cuando:**
   - Se alcanza un número predefinido de iteraciones o
   - La diferencia entre las imputaciones sucesivas es suficientemente pequeña.

Teóricamente para variables cuantitativas tenemos:

Asuma que $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$ es una matriz de datos $n \times p$. Para una variable arbitraria $\mathbf{X}_s$ que incluye valores faltantes en las posiciones $\mathbf{i}_{mis}^{(s)} \subseteq \{1, \ldots, n\}$, con $s = 1, \ldots, p$, podemos separar el conjunto de datos en 4 partes:

- Los valores observados de la variable $\mathbf{X}_s$, denotado por $\mathbf{y}_{obs}^{(s)}$.
- Los valores faltantes de la variable $\mathbf{X}_s$, denotado por $\mathbf{y}_{mis}^{(s)}$.
- La variables distintas a $\mathbf{X}_s$ con observaciones $\mathbf{i}_{obs}^{(s)} = \{1, \ldots, n\} \setminus \mathbf{i}_{mis}^{(s)}$ denotado por $\mathbf{x}_{obs}^{(s)}$.
- Las variables distintas a $\mathbf{X}_s$ con observaciones $\mathbf{i}_{mis}^{(s)}$ denotado por $\mathbf{x}_{mis}^{(s)}$.

Por tanto, el algoritmo de MissForest empleando Random Forest será el mostrado en la figura 3.

Algorithm 1 Impute missing values with RF.

Require: $\mathbf{X}$ an $n \times p$ matrix, stopping criterion γ

1. Make initial guess for missing values;
2. $\mathbf{k} \leftarrow$ vector of sorted indices of columns in $\mathbf{X}$ w.r.t. increasing amount of missing values;
3. **while** not γ **do**
4. $\quad \mathbf{X}_{\text{old}}^{\text{imp}} \leftarrow$ store previously imputed matrix;
5. $\quad$ **for** s in $\mathbf{k}$ **do**
6. $\quad\quad$ Fit a random forest: $\mathbf{y}_{\text{obs}}^{(s)} \sim \mathbf{x}_{\text{obs}}^{(s)}$;
7. $\quad\quad$ Predict $\mathbf{y}_{\text{mis}}^{(s)}$ using $\mathbf{x}_{\text{mis}}^{(s)}$;
8. $\quad\quad \mathbf{X}_{\text{new}}^{\text{imp}} \leftarrow$ update imputed matrix, using predicted $\mathbf{y}_{\text{mis}}^{(s)}$;
9. $\quad$ **end for**
10. $\quad$ update γ.
11. **end while**
12. **return** the imputed matrix $\mathbf{X}^{\text{imp}}$

Figura 3: Algoritmo de imputación de valores faltantes por RF Stekhoven and Bühlmann [2011].

Y teóricamente para variables mixtas tenemos:

Para la comparación entre diferentes tipos de variables, podemos aplicar el algoritmo KKNimpute (Faisal and Tutz [2022], Stekhoven and Bühlmann [2011]) con variables dummy para las variables categóricas. Esto se hace codificando una variable categórica $\mathbf{X}_j$ en m variables dicotómicas $\widetilde{\mathbf{X}}_{j,m} \in \{-1, 1\}$. La aplicación del algoritmo KKNimpute para variables categóricas lo podemos resumir como:

- Codificar todas las variables categóricas en $\{-1, 1\}$ como variables dummy.
- Estandarizar todas las variables a media 0 y desviación estándar 1.
- Aplicar el método de validación cruzada KKNimpute de la figura 4.
- Retransformar la matriz de datos imputada a las escalas originales.
- Volver a codificar las variables dummy a las variables categóricas originales.
- Calcular el error de imputación.

Algorithm 2 Cross-validation KNN imputation.

Require: $\mathbf{X}$ an $n \times p$ matrix, number of validation sets l, range of suitable number of nearest neighbours $\mathbf{K}$

1. $\mathbf{X}^{\text{CV}} \leftarrow$ initial imputation using mean imputation;
2. **for** t in $1, \ldots, l$ **do**
3. $\quad \mathbf{X}^{\text{CV}}_{\text{mis},t} \leftarrow$ artificially introduce missing values to $\mathbf{X}^{\text{CV}}$;
4. $\quad$ **for** k in $\mathbf{K}$ **do**
5. $\quad\quad \mathbf{X}^{\text{CV}}_{\text{KNN},t} \leftarrow$ KNN imputation of $\mathbf{X}^{\text{CV}}_{\text{mis},t}$ using k nearest neighbours;
6. $\quad\quad \varepsilon_{k,t} \leftarrow$ error of KNN imputation for k and t;
7. $\quad$ **end for**
8. **end for**
9. $k_{\text{best}} \leftarrow \underset{k}{\text{argmin}} \frac{1}{l} \sum_{t=1}^{l} \varepsilon_{k,t}$;
10. $\mathbf{X}^{\text{imp}} \leftarrow$ KNN imputation of $\mathbf{X}$ using k_{best} nearest neighbours.

Figura 4: Algoritmo de imputación de valores faltantes por KNNimpute Stekhoven and Bühlmann [2011].

Métricas de Evaluación

Para evaluar el desempeño de MissForest, se utilizan dos métricas dependiendo de los tipos de variables que se encuentran en el conjunto de datos:

Raiz del Error Cuadrático Medio Normalizado (NRMSE)

$$\text{NRMSE} = \sqrt{\frac{\text{mean}\left(\left(\mathbf{X}^{\text{true}} - \mathbf{X}^{\text{imp}}\right)^2\right)}{\text{var}\left(\mathbf{X}^{\text{true}}\right)}}$$

Esta métrica es una versión normalizada del Root Mean Squared Error (RMSE) y proporciona una medida relativa del error en relación con la escala de los datos cuantitativos (Oba et al. [2003]).

Proporción media de categorías imputadas falsamente (PFC)

$$\Delta F = \frac{\sum_{j \in \mathbf{F}} \sum_{i=1}^{n} \mathbf{I}_{\mathbf{X}_{\text{new}}^{\text{imp}} \neq \mathbf{X}_{\text{old}}^{\text{imp}}}}{\#\text{NA}}$$

Donde $\mathbf{F}$ es el conjunto de las variables categóricas y #NA es el número de valores faltantes en las variables categóricas. Evalúa la proporción de valores imputados que coinciden con los valores observados (Faisal and Tutz [2022]).

Tipos de Variables

MissForest (Stekhoven and Bühlmann [2011]) es versátil y puede manejar conjuntos de datos que contienen tanto variables categóricas como numéricas. Además, se puede aplicar a diferentes tipos de variables:

- **Variables Numéricas Continuas:** Efectivo para imputar valores faltantes en variables numéricas continuas.
- **Variables Categóricas:** Puede manejar variables categóricas al incluirlas en el proceso de imputación de bosques aleatorios.
- **Datos Mixtos:** Aplicable a conjuntos de datos con una combinación de variables numéricas y categóricas.
- **Escalas Diferentes:** Adaptado a datos con diferentes escalas, ya que los bosques aleatorios no son sensibles a la escala de las variables.
- **Datos Complejos:** Funciona bien en conjuntos de datos complejos con patrones no lineales y múltiples variables con datos faltantes.

Espacio de Hilbert separable

Por Ferraty [2006] una variable aleatoria X_k es una observación o variable funcional si toma valores en un espacio funcional H (Espacio normado o semi-normado completo). Se define que H es un espacio de Hilbert separable (es decir que tiene una base contable $e_k, k \in \mathbb{Z}$), donde $\{e_i\}$ es una base ortogonal arbitraria, con producto interior $\langle \cdot, \cdot \rangle$ con la norma $\| \cdot \|$ generada. Todas las funciones aleatorias son definidas en algún espacio de probabilidad común $(\Omega, \mathcal{F}, P)$.

Uno de los ejemplos más importantes de espacio de Hilbert es $L^2 = L^2([0,1])$, el cual es muy utilizado. Todas las funciones aleatorias son definidas en algún espacio de probabilidad común $(\Omega, \mathcal{F}, P)$.

Se dice que la función aleatoria X es integrable Ferraty [2006] si $E\|X\| < \infty$, y cuadrado integrable si $E\|X\|^2 < \infty$. Para el caso de $E\|X\|^p < \infty$, $p > 0$, se escribe $X \in L_H^p = L_H^p(\Omega, \mathcal{F}, P)$. Se habla de convergencia en media de X_n a X en L_H^p si $E\|X_n - X\|^p \to 0$ mientras que $\|X_n - X\| \to 0$ hace referencia a la convergencia casi siempre.

Operadores funcionales

Sea $\mathcal{L}$ el espacio de los operadores lineales acotados continuos sobre H con la norma (Ferraty [2006]):

$$\|\Psi\|_{\mathcal{L}} = sup\left\{\|\Psi(x)\| : \|x\| \leq 1\right\}$$

Un operador $\Psi \in \mathcal{L}$ es compacto si existen dos bases ortonormales v_j y f_j, y una sucesión de valores reales λ_j asumidos positivos que converge a cero, tal que:

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle f_j, \ x \in H \tag{1}$$

Se asume que los valores de λ son positivos. La representación de la ecuación (1) es llamada descomposición del valor singular (Ferraty [2006]).

Si el operador admite representarse como la ecuación (1) y $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$, entonces es un operador *Hilbert - Schmidt* (Ferraty [2006]).

El espacio S de operadores de *Hilbert - Schmidt* es un espacio de Hilbert separable con un producto escalar:

$$\langle \Psi_1, \Psi_2 \rangle_S = \sum_{i=1}^{\infty} \langle \Psi_1(e_i), \Psi_2(e_i) \rangle$$

donde $\{e_i\}$ es una base ortonormal arbitraria. Esto implica que $\|\Psi\|_S^2 = \sum_{j \geq 1} \lambda_j^2$ y $\|\Psi\|_{\mathcal{L}} \geq \|\Psi\|_S$.

Un operador $\Psi \in \mathcal{L}$ es simétrico si (Ferraty [2006]):

$$\langle \Psi(x), y \rangle = \langle x, \Psi(y) \rangle, \ x, y \in H$$

y es definido positivo si

$$\langle x, \Psi(x) \rangle > 0, \ x \in H \backslash \{0\}$$

Un operador Ψ de *Hilbert - Schmidt* simétrico definido positivo admite una descomposición

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j, \ x \in H$$

con v_j bases ortonormales, que son funciones propias de Ψ, es decir, $\Psi(v_j) = \lambda_j v_j$. Los v_j pueden ser asumidos como una base.

Espacio L^2

El espacio L^2 es un conjunto medible de funciones de valor real x en el intervalo $[0, 1]$ que satisfacen $\int_0^1 x^2(t)dt < \infty$. Este es un espacio de Hilbert separable con producto interior (Ferraty [2006]):

$$\langle x, y \rangle = \int x(t)y(t)dt. \tag{2}$$

Una importante clase de operadores en L^2 son los operadores integrales definidos por

$$\Psi(x)(t) = \int \psi(t, s)x(s)ds, \quad x \in L^2 \tag{3}$$

con $\psi(\cdot, \cdot)$ kernel real. Tal operador es $Hilbert - Schmidt$ si y sólo si $\iint \psi^2(t, s)dtds < \infty$, en cuyo caso

$$\|\Psi\|_S^2 = \iint \psi^2(t,s)dtds. \tag{4}$$

Si $\psi(s,t) = \psi(t,s)$ y $\iint \psi(t,s)x(t)x(s)dtds \geq 0$, el operador integral Ψ es simétrico y definido positivo, lo cual permite que

$$\psi(t,s) = \sum_{j=1}^{\infty} \lambda_j v_j(t) v_j(s) \quad para \quad L^2([0,1]\times[0,1]). \tag{5}$$

Si ψ es continua, la expansión en series mostrada arriba se cumple para toda s, $t \in [0,1]$, y la serie es uniformemente convergente. Este resultado es conocido como el teorema de Mercer (Riesz [1990]).

Operador de media funcional

Sean $X_1, X_2, X_3, \ldots$ funciones aleatorias en H. Se llama a X débilmente integrable si existe $\mu \in H$ tal que $E\langle X, y\rangle = \langle \mu, y\rangle$, para todo $y \in H$. En este caso μ es el valor esperado de X, de manera corta EX (Ramsay [2006]). Algunos resultados elementales son:

- EX es única.
- Integrabilidad implica integrabilidad débil.
- $\|EX\| \leq E\|X\|$.

Si $H = L^2$ se puede mostrar que

$$\{(EX)(t), t \in [0,1]\} = \{E(X(t)), t \in [0,1]\}$$

lo que indica que se puede obtener el valor esperado evaluado en cada punto y EX conmuta con operadores acotados, es decir, si $\Psi \in \mathcal{L}$ y X es integrable, entonces:

$$E\Psi(X) = \Psi(EX)$$

Operador de covarianza funcional

Para $X \in L_H^2$ el operador covarianza de X está definido por:

$$C(y) = E\left[\langle X - EX, y\rangle (X - EX)\right], \; y \in H$$

El operador C es simétrico y definido positivo, con funciones propias λ_i que satisfacen:

$$\sum_{i=1}^{\infty} \lambda_i = E\|X - E(X)\|^2 < \infty$$

Por tanto, $C \in S$, es definido positivo y simétrico, lo que permite, por el teorema de Mercer (Riesz [1990]), representarlo como en (5).

Estimador de operadores muestrales de media y covarianza

El estimador de los operadores muestrales de media y covarianza de funciones aleatorias $X_1, X_2, \ldots, X_N$ están definidos como (Ferraty [2006]):

$$\hat{\mu}_N = \frac{1}{N}\sum_{k=1}^{N} X_k$$

$$\hat{C}_N(y) = \frac{1}{N}\sum_{k=1}^{N}\langle X_k - \hat{\mu}_N, y\rangle\,(X_k - \hat{\mu}_N)\,,\ y \in H$$

El siguiente resultado implica la consistencia de los estimadores definidos para muestras idénticamente distribuidas.

Teorema 1: Sea $\{X_n\}$ una sucesión de variables funcionales aleatorias en H idénticamente distribuidas, con $EX = \mu$ (Ferraty [2006]).

- Si $X_1 \in L_H^2$ entonces $E\|\hat{\mu}_N - \mu\|^2 = O(N^{-1})$.
- Si $X_1 \in L_H^4$ entonces $E\|\hat{C}\|_S^2 < \infty$ y $E\|C - \hat{C}\|_S^2 = O(N^{-1})$.

De $H = L^2$ se utiliza:

$$C(y)(t) = \int c(t,s)y(s)ds$$

donde $c(t,s) = Cov(X(t), X(s))$. El núcleo de la covarianza es estimado por:

$$\hat{c}(t,s) = \frac{1}{N}\sum_{k=1}^{N}(X_k(t) - \hat{\mu}_N(t))(X_k(s) - \hat{\mu}_N(s))$$

Bases de funciones

Dada la definición por Ferraty [2006], una base es un conjunto de funciones conocidas e independientes $\{u_k\}_{k\in\mathbb{N}}$ tales que cualquier función puede ser aproximada, tan bien como se quiera, mediante una combinación lineal de ellas con k suficientemente grande. De esta forma, la observación funcional $X_n(t)$ puede representarse como

$$X_n(t) \approx \sum_{k=1}^{K}\lambda_k u_k(t) \tag{6}$$

Si el dato funcional pertenece al espacio de Hilbert, esto garantiza que existe una base ortonormal tal que $X_n(t) = \sum_{k=1}^{\infty}\langle X, v_k\rangle v_k$. Ahora teniendo el producto interno, se pueden estimar los coeficientes resolviendo el

siguiente sistema:

$$\begin{pmatrix} \langle X, u_1 \rangle \\ . \\ . \\ . \\ \langle X, u_i \rangle \\ . \\ . \\ . \\ \langle X, u_k \rangle \end{pmatrix} = \begin{pmatrix} \langle u_1, u_1 \rangle & . & . & . & \langle u_i, u_1 \rangle & . & . & . & \langle u_k, u_1 \rangle \\ . & & & & . & & & & . \\ . & & & & . & & & & . \\ . & & & & . & & & & . \\ \langle u_1, u_i \rangle & . & . & . & \langle u_i, u_i \rangle & . & . & . & \langle u_k, u_i \rangle \\ . & & & & . & & & & . \\ . & & & & . & & & & . \\ . & & & & . & & & & . \\ \langle u_1, u_k \rangle & . & . & . & \langle u_i, u_k \rangle & . & . & . & \langle u_k, u_k \rangle \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 \\ . \\ . \\ . \\ \lambda_i \\ . \\ . \\ . \\ \lambda_k \end{pmatrix} \tag{7}$$

B-splines

Esta representación será especialmente útil para funciones estables, sin grandes variaciones y con una curvatura más o menos constante, los B-spline o spline cúbicos son una buena base para aproximar las funciones y una de la más utilizada (Ferraty [2006]). Dividir el subintervalo $[0, T]$ en L subintervalos separados por los puntos $a = t_0, t_1, ..., t_L = b$. en cada uno de estos intervalos el spline es un polinomio de cierto orden m que ajusta la curva, y a su vez estos polinomios generan una base ortogonal.

Por Ferraty [2006] tenemos que sea $B_k(t, \tau)$ el valor del k-ésimo elemento de la base sobre una partición τ en el instante t, la función spline $S(t)$ está definida como:

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau)$$

Las llamadas bases de B-splines permiten aproximar todas estas funciones facilitando así su manejo. Sus características se basan en las siguiente propiedades:

1. Cada elemento de la base $u_k(t)$ será una función spline de orden m y partición τ.
2. Cualquier combinación lineal de funciones spline es una función spline.
3. Cualquier función spline de orden m sobre la partición τ se puede expresar como combinación lineal de las funciones de la base.

Análisis de Componentes Principales Funcionales

La idea principal de los componentes principales funcionales (CPF) es caracterizar una función mediante un conjunto de funciones ortonormales, de forma que la proyección de dicha función en este conjunto simule lo mejor posible el comportamiento de dicha función. Todo esto se realiza a partir de las funciones propias de manera similar al realizado en el análisis multivariado con los vectores y valores propios (Ramsay [2006]).

Asuma que un conjunto de realizaciones funcionales aleatorias $x_1(t), x_2(t), \ldots, x_N(t)$ es un espacio de Hilbert separable. Suponga que los datos están centrados, es decir que $\sum_{i=1}^{N} x_i = 0$. Fijando un entero $l < N$. Se desea encontrar una base ortonormal $u_1, u_2, \ldots, u_p$ tal que:

$$\hat{S}^2 = \sum_{i=1}^{N} \|x_i - \sum_{k=1}^{l} \langle x_i, x_k \rangle u_k\|^2$$

es mínimo. Si la base u_j existe, entonces x_i es aproximadamente $\sum_{k=1}^{l} \langle x_i, u_k \rangle u_k$, para algún valor l, es decir, en lugar de trabajar con curvas de dimensión infinita x_i se trabajará con vectores finitos l-dimensionales:

$$\mathbf{x}_i = [\langle x_i, u_1\rangle, \langle x_i, u_2\rangle, \ldots, \langle x_i, u_l\rangle]^T$$

Las funciones $u_1, u_2, \ldots, u_l$ que minimizan a $\hat{S}^2$ son las funciones propias del operador de covarianza muestral, es decir, $\hat{C}(u_i) = \hat{\lambda}_i u_i$, donde $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_l$.

A las funciones con esta particularidad se les llama componentes principales empíricos (CPEF) de los datos $x_1(t), x_2(t), \ldots, x_N(t)$.

En la mayoría de ocasiones se estiman los valores propios y las funciones propias de C. Los valores propios deben ser identificables, por tanto se debe asumir que $\lambda_1 > \lambda_2 > \ldots$, en la practica, solo se pueden estimar a lo más l valores propios y se asume que $\lambda_1 > \lambda_2 > \ldots > \lambda_{l-1} > \lambda_l$, lo cual implica que los primeros l valores propios son distintos de cero. Por otra parte, las funciones propias son definidas por $C(v_j) = \lambda_j v_j$, esto es v_j es una función propia.

Una forma de estimar los valores propios as a través de Bosq [2012]:

$$\hat{C}(\hat{v}_j) = \hat{\lambda}_j \hat{v}_j, \;\; j = 1, 2, \ldots, N$$

Teorema 2: (Bosq [2012]) Se asume que las observaciones $X, X_1, X_2, \ldots, X_N$ son *i.i.d.* en H tal que L_H^4 con $EX = 0$. Se supone que:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_d > \lambda_{d+1}$$

Entonces, para cada $1 \leq j \leq d$,

$$E\left[\|\hat{c}_j \hat{v}_j - v_j\|^2\right] = O\left(N^{-1}\right)$$

y

$$E\left[\mid \lambda_j - \hat{\lambda}_j \mid^2\right] = O\left(N^{-1}\right)$$

El teorema implica que, bajo condiciones de regularidad, la población de funciones propias puede tener como estimador consistente al conjunto de funciones propias empíricas (Bosq [2012]).

Dados C y K dos operadores compactor en $\mathcal{L}$, se define operadores con valores singulares de descomposición si:

$$C(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j\rangle f_j, \quad K(x) = \sum_{j=1}^{\infty} \gamma_j \langle x, u_j\rangle g_j \tag{8}$$

donde λ_j y γ_j son valores propios y $v_j,\ f_j,\ u_j$, y g_j funciones propias, correspondientes a C y K.

Lema 1: Sean $C,\ K \in \mathcal{L}$ dos operadores que satisfacen 8, entonces para cada $j \geq 1$, $\mid \gamma_j - \lambda_j \mid \leq \|K - C\|_{\mathcal{L}}$.

Lema 2: Sean $C,\ K \in \mathcal{L}$ dos operadores que satisfacen 8. Si C es simétrico y además $f_j = v_j$, entonces:

$$\|u_j - v_j'\| \leq \frac{2\sqrt{2}}{\alpha_j}\|K - C\|_{\mathcal{L}}, \quad 1 \leq j \leq d$$

donde $\alpha_1 = \lambda_1 - \lambda_2$ y $\alpha_j = min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$, con $2 \leq j \leq d$.

Para el caso particular en el cual C es el operador de covarianza, entonces este satisface las condiciones impuestas en el Lema 2. Los v_j son las funciones propias de C (Ramsay [2006]).

Proceso espacial autorregresivo hilbertiano ($SARH(1)$)

Sea H un espacio de Hilbert separable de funciones sobre un dominio $D \subseteq \mathbb{R}^n$ Bosq [2012], Ruiz-Medina [2012], Ruiz and Angulo [2007], con producto interno $\langle \cdot, \cdot \rangle_H$, y con norma denotada por $\| \cdot \|_H$. Sea $\{X_t, t \in \mathbb{N}\}$, un proceso estacionario de Hilbert sobre un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ satisface

$$X_t(s) = \Psi X_{t-1}(s) + v_t(s), \quad s \in D \subseteq \mathbb{R}^n, t \in \mathbb{N}, \tag{9}$$

donde v_t es RBF funcional, es decir, una sucesión de variables aleatorias independientes e idénticamente distribuidas sobre H y que satisface

$$E[\|v_n\|_H^2] = \sigma_v^2 < \infty \tag{10}$$

El operador de autocorrelación Ψ en (9) es acotado y definido sobre un dominio denso en H. Para el caso en que Ψ está definido mediante una expresión integral en términos de un kernel homogéneo y estable ante convoluciones, el modelo (9) permite una aproximación de los modelos definidos en espacio y tiempo continuos. Por otro lado, Ψ admite descomposición espectral mediante el patrón de oscilaciones principales (POP) definido a continuación.

Sea Ψ el operador de autocorrelación en el modelo (9), representado por la interacción espacio temporal en D. El operador Ψ satisface las siguientes ecuaciones

$$\begin{aligned} &\Psi\gamma_i = \lambda_i\gamma_i, \quad i \in \mathbb{N}, \\ &\text{y} \\ &\Psi^*\phi_i = \lambda_i\phi_i, \quad i \in \mathbb{N}, \end{aligned} \tag{11}$$

donde γ_i y ϕ_i son los sistemas de funciones propias asociados al espectro puntual $\{\lambda_i, i \in \mathbb{N}\}$ de los operadores Ψ y Ψ^*, es decir un operador no simétrico.

Inferencia por medio de $SARH(1)$

Algoritmo EM en el modelo SARH(1)

En este apartado se muestra cómo usar el algoritmo EM, para hacer estimación del operador del modelo (9).

Los valores iniciales que se proponen en el algoritmo son los estimadores de los operadores de covarianza y covarianza cruzada R_{Z_0} y $R_{X_0X_1} = R_{Z_0Z_1}$ obtenidos a partir del método de los momentos. Toda la deducción matemática del algoritmo es desarrollada en Bosq [2012].

Si el operador de covarianza de ruido es proporcional a la identidad, el estimador inicial $\hat{\sigma}_\epsilon$ que define la diagonal principal del operador se obtiene por los datos, teniendo en cuenta el efecto *Pepita* del variograma empírico en una vecindad alrededor a cero. El operador de covarianza espacial se estima como

$$\hat{R}_{X_0} = \hat{R}_{Z_0} - \hat{\sigma}_\epsilon. \tag{12}$$

Así mismo los operadores de covarianza espacial y covarianza cruzada se definen mediante las ecuaciones

$$R_{X_0}(\phi) = E[X_0\langle X_0, \phi\rangle_H], \quad \forall \phi \in D(R_{X_0}), \tag{13}$$
$$R_{X_0X_1}(\phi) = E[X_0\langle X_1, \phi\rangle_H], \quad \forall \phi \in D(R_{X_0X_1}), \tag{14}$$

donde D es el dominio de funciones de los operadores de covarianza. Las estimaciones de momentos de los operadores de covarianza y covarianza cruzada están dados por

$$\hat{R}_{X_0} = \frac{1}{T}\left[\sum_{i=1}^{T} Z_i \otimes Z_i\right] - \left[\frac{1}{T}\sum_{i=1}^{T} Z_i\right] \otimes \left[\frac{1}{T}\sum_{i=1}^{T} Z_i\right] - \hat{\sigma}_\epsilon^2 I, \tag{15}$$
$$\hat{R}_{X_0X_1} = \frac{1}{T-1}\left[\sum_{i=1}^{T-1} X_i \otimes X_{i+1}\right] - \left[\frac{1}{T-1}\sum_{i=1}^{T-1} X_i\right] \otimes \left[\frac{1}{T-1}\sum_{i=1}^{T-1} X_{i+1}\right], \tag{16}$$

con $\hat{\sigma}_\epsilon^2$ el estimador inicial basado en el efecto *pepita* del semivariograma empírico funcional

$$\hat{\gamma}(\triangle t) = \frac{1}{2N(\triangle t)} \sum_{i,j \in S(\triangle t)} \|X_{t_i} - X_{t_j}\|_H^2, \tag{17}$$

para $S(\triangle t) = \{(i,j) | |t_i - t_j| = \triangle t\}$, y $N(\triangle t)$ el número de pares en $S(\triangle t)$.

El estimador funcional (17), tiene propiedades similares al estimador clásico del variograma. A partir de las ecuaciones (15) y (16) se define las estimaciones del operador diagonal Λ asociado al espectro puntual del operador de autocorrelación

$$\hat{\Lambda} = \Phi^* \hat{R}_{X_0X_1} \hat{R}_{X_0}^{-1} \Gamma \tag{18}$$
$$y$$
$$\Phi^* \hat{R}_\epsilon \Phi = \Phi^* \hat{R}_{X_0} \Phi - \Phi^* \hat{R}_{X_0X_1} \Phi \hat{\Lambda}^T - \hat{\Lambda} \Phi^* \hat{R} *_{X_0X_1} \Phi + \hat{\Lambda} \Phi^* \hat{R}_{X_0} \Phi \hat{\Lambda}^T \tag{19}$$

donde Φ y Γ son los operadores de proyección sobre los subespacios generados por el conjunto de funciones propias. Cuando estas se conocen, se pueden utilizar la diagonalización POP del estimador Ψ obtenido mediante la ecuación

$$\hat{\Psi} = \hat{R}_{X_0X_1} \hat{R}_{X_0}^{-1}. \tag{20}$$

Existen diferentes métodos numéricos para realizar la inversa del operador Bosq [2012]. El estimador $\hat{\Psi}^T = \hat{R}_{X_0X_1}^T [\hat{R}_{X_0}^T]^{-1}$ se proyecta en las observaciones funcionales espaciales para cada uno de los T tiempos de observación sobre un subespacio de dimensión finita, obteniendo

$$\hat{\Psi}^T = \hat{\Phi}_l \hat{R}_{X_0X_1}^T (\hat{R}_{X_0}^{T,l})^{-1} \hat{\Phi}_l, \tag{21}$$

donde $\hat{\Phi}_l$ es el operador de proyección sobre el subespacio de H generado por l funciones propias empíricas y

$$\hat{R}_{X_0}^{T,M} = \sum_{j=1}^{M} \hat{\lambda}_j \hat{\phi}_j \otimes \hat{\phi}_j, \tag{22}$$

siendo $\{\hat{\lambda}_j\}_{j\in\mathbb{N}}$ los valores propios asociados al sistema de funciones propias empíricas $\{\hat{\phi}_j\}_{j\in\mathbb{N}}$ Bosq [2012] y l es el número de funciones óptimas para el operador de proyección $\hat{\Psi}^T$.

Si el operador de covarianza del ruido de observación no es diagonal, la proyección del operador R_ϵ, sobre el conjunto de funciones propias de R_{X_0} está dado por

$$\Phi^* R_\epsilon = \Phi^* R_{Z_0} - \hat{\Lambda}\Phi^*. \tag{23}$$

Dado que

$$\Phi^* R_{X_0} = \hat{\Lambda}\Phi^*, \tag{24}$$

con $R_{X_0} = \Phi\hat{\Lambda}\Phi^*$, $\Phi\Phi^* = I$ y $\hat{\Lambda}$ el operador diagonal definido a partir del espectro puntual de R_{X_0}, reemplazando en (23) los operadores Φ y Λ se obtiene una estimación del operador R_ϵ.

El implementación del algoritmo EM en el modelo ARH(1), bajo el siguiente supuesto

$$\Phi_l^*(X_t(\cdot) - \Psi X_{t-1}(\cdot)) \sim N(0, \Phi_l^* R_v \Phi_l). \tag{25}$$

Puesto que

$$\Phi_l^*(X_t(\cdot) - \Psi X_{t-1}(\cdot)) = a(t) - \Lambda a(t-1), \tag{26}$$

donde $J(t) = a(t) - \Lambda_l a(t-1)$, se tiene

$$f_J(j; \Psi, R_v) = \frac{1}{\sqrt{2\pi\Phi_l^* R_v \Phi_l}} \cdot \exp\left\{-\frac{J(t)^T(\Phi_l^* R_v \Phi)^{-1} J(t)}{2}\right\}. \tag{27}$$

De manera similar bajo el supuesto que

$$e(t) \sim N(0, \Phi_l^* R_\epsilon \Phi_l), \tag{28}$$

se tiene

$$f_\epsilon(e; R_\epsilon) = \frac{1}{\sqrt{2\pi\Phi_l^* R_\epsilon \Phi_l}} \cdot \exp\left\{-\frac{e(t)^T(\Phi_l^* R_\epsilon \Phi)^{-1} e(t)}{2}\right\}, \tag{29}$$

donde $e(t) = \Phi_l^* \epsilon_t$.

Dada una muestra aleatoria,

$$\begin{aligned} f_J(J_1(\cdot), ..., J_T(\cdot); \Psi, R_v) &= \frac{1}{(\sqrt{2\pi\Phi_l^* R_v \Phi_l})^T} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{T} J(i)^T(\Phi_l^* R_v \Phi)^{-1} J(i)\right\}. \\ f_\epsilon(e_1(\cdot), ..., e_T(\cdot); R_\epsilon) &= \frac{1}{(\sqrt{2\pi\Phi_l^* R_\epsilon \Phi_l})^T} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{T} e(i)^T(\Phi_l^* R_v \Phi)^{-1} e(i)\right\}. \end{aligned} \tag{30}$$

Dado que ϵ es independiente de X,

$$f_{J,\epsilon} = f(j; \Theta_1) \cdot f(\epsilon; \Theta_2), \tag{31}$$

donde $\Theta = ((\Psi, R_v), R_\epsilon)^T = (\Theta_1, \Theta_2)$ los parámetros funcionales que definen las densidades de v y ϵ. De (31) se obtiene

$$\begin{aligned}\ln(f_{J,\epsilon}) &= \ln(f(j;\Theta_1)\cdot f(\epsilon;\Theta_2)) \\ &= C - \frac{T}{2}[\ln|\Phi_l^* R_v \Phi_l| + \ln|\Phi_l^* R_\epsilon \Phi_l|] \\ &\quad - \frac{1}{2}\sum_{i=1}^{T} J(i)^T (\Phi_l^* R_v \Phi_l)^{-1} J(i) \\ &\quad - \frac{1}{2}\sum_{i=1}^{T} e(i)^T (\Phi_l^* R_\epsilon \Phi_l)^{-1} e(i), \end{aligned} \tag{32}$$

con C una constante que no depende del vector de parámetros funcionales. Se verifica entonces que

$$\begin{aligned}\sum_{i=1}^{T} J(i)^T (\Phi_l^* R_v \Phi_l)^{-1} J(i) &= traza\left(\sum_{i=1}^{T} J(i)^T (\Phi_l^* R_v \Phi_l)^{-1} J(i)\right) \\ &= traza\left(\sum_{i=1}^{T} (\Phi_l^* R_v \Phi_l)^{-1} J(i) J(i)^T\right) \\ &= traza\left((\Phi_l^* R_v \Phi_l)^{-1} \sum_{i=1}^{T} J(i) J(i)^T\right), \end{aligned} \tag{33}$$

y

$$\begin{aligned} J(i)J(i)^T &= (a(i) - \Lambda_l a(i-1))(a(i) - \Lambda_l a(i-1))^T \\ &= a(i)a(i)^T - a(i)a(i-1)^T\Lambda_l^T - \Lambda_l a(i-1)a(i)^T + \Lambda_l a(i-1)a(i-1)^T\Lambda_l^T, \end{aligned} \tag{34}$$

se obtiene

$$traza\left((\Phi_l^* R_v \Phi_l)^{-1}\left[\sum_{i=1}^{T}(a(i)a(i)^T - a(i)a(i-1)^T\Lambda_l^T - \Lambda_l a(i-1)a(i)^T + \Lambda_l a(i-1)a(i-1)^T\Lambda_l^T)\right]\right) \tag{35}$$

Análogamente para $e(t)$

$$\sum_{i=1}^{T} e(i)^T (\Phi_l^* R_\epsilon \Phi_l)^{-1} e(i) = traza\left((\Phi_l^* R_v \Phi_l)^{-1} \sum_{i=1}^{T} e(i)e(i)^T\right). \tag{36}$$

Por tanto (32) se escribe como

$$\ln(f_{J,\epsilon}) = C - \frac{T}{2}[\ln|\Phi_l^* R_v \Phi_l| + \ln|\Phi_l^* R_\epsilon \Phi_l|] - \frac{1}{2} traza\left((\Phi_l^* R_v \Phi_l)^{-1} \left[\sum_{i=1}^{T} \left(a(i)a(i)^T - a(i)a(i-1)^T \Lambda_l^T - \Lambda_l a(i-1)a(i)^T + \Lambda_l a(i-1)a(i-1)^T \Lambda_l^T \right) \right] \right) - \frac{1}{2} traza\left((\Phi_l^* R_\epsilon \Phi_l)^{-1} \left[\sum_{i=1}^{T} e(i)e(i)^T \right] \right). \tag{37}$$

Se tiene

$$E_{\hat{\Theta}(i)}[\ln f_{J,\epsilon}(j, e : \Theta) \mid Z] = C - \frac{T}{2}\left[\ln|\Phi_l^* R_v \Phi_l| + \ln|\Phi_l^* R_\epsilon \Phi_l| \right] - \frac{1}{2} traza\left((\Phi_l^* R_v \Phi_l)^{-1} \left[C_X - B_X \Lambda_l^T - \Lambda_l B_X^T + \Lambda_l A_X \Lambda_l^T \right] \right) - \frac{1}{2} traza\left((\Phi_l^* R_v \Phi_l)^{-1} C_\epsilon \right), \tag{38}$$

donde $\Theta(i)$ representa la estimación de Θ en la iteración i y

$$C_X = \sum_{i=1}^{T} E[a(i)a(i)^T \mid Z], \tag{39}$$

$$B_X = \sum_{i=1}^{T} E[a(i)a(i-1)^T \mid Z], \tag{40}$$

$$A_X = \sum_{i=1}^{T} E[a(i-1)a(i-1)^T \mid Z], \tag{41}$$

$$C_\epsilon = \sum_{i=1}^{T} E[e(i)e(i)^T \mid Z]. \tag{42}$$

$$\tag{43}$$

Realizando la derivada parcial respecto a $\Phi_l^* R_v \Phi$ en (38)

$$\frac{\partial E_{\hat{\Theta}(i)}[\ln f_{J,\epsilon}(j, e : \Theta) \mid Z]}{\partial(\Phi_l^* R_v \Phi_l)} = \frac{\partial}{\partial(\Phi_l^* R_v \Phi_l)} \left(-\frac{T}{2} \ln|\Phi_l^* R_v \Phi_l| - \frac{1}{2} traza\left((\Phi_l^* R_v \Phi_l)^{-1} D_X \right) \right) = \frac{T}{2}(\Phi_l^* R_v \Phi_l)^{-1} - \frac{1}{2}\left((\Phi_l^* R_v \Phi_l)^{-1} D_X (\Phi_l^* R_v \Phi_l)^{-1} \right) \tag{44}$$

con $D_X = C_X - B_X \Lambda_l^T - \Lambda_l B_x^T + \Lambda_l A_X \Lambda_l^T$.

E igualando a cero la ecuación anterior se obtiene

$$0 = \frac{T}{2}(\Phi_l^* R_v \Phi_l)^{-1} - \frac{1}{2}\left(-(\Phi_l^* R_v \Phi_l)^{-1} D_X (\Phi_l^* R_v \Phi_l)^{-1}\right). \tag{45}$$

despejando $\Phi_l^* R_v \Phi_l$ de arriba se llega a

$$\widehat{\Phi_l^* R_v \Phi_l} = \frac{C_X - B_X \hat{\Lambda}_l^T - \hat{\Lambda}_l B_x^T + \hat{\Lambda}_l A_X \hat{\Lambda}_l^T}{T}. \tag{46}$$

Análogamente, realizando la derivada parcial de (38) respecto a $\Phi_l^* R_\epsilon \Phi$

$$\widehat{\Phi_l^* R_\epsilon \Phi_l} = \frac{C_\epsilon}{T}. \tag{47}$$

Ahora derivando respecto a Λ

$$\frac{\partial E_{\hat{\Theta}(i)}[\ln f_{J,\epsilon}(j, e : \Theta) \mid Z]}{\partial \Lambda} = \frac{\partial}{\partial \Lambda}\left(traza(\Phi_l^* R_v \Phi_l)^{-1} D_X\right) \tag{48}$$

$$= -(\Phi_l^* R_v \Phi_l)^{-1} B_X - [(\Phi_l^* R_v \Phi_l)^{-1}]^T B_X + [(\Phi_l^* R_v \Phi_l)^{-1}]^T \Lambda_l A_X^T \tag{49}$$

$$+ (\Phi_l^* R_v \Phi_l)^{-1} \Lambda_l A_X, \tag{50}$$

puesto que $\Phi_l^* R_v \Phi_l$ y A_X son simétricas

$$\frac{\partial E_{\hat{\Theta}(i)}[ln f_{J,\epsilon}(j, e : \Theta) \mid Z]}{\partial \Lambda} = -2(\Phi_l^* R_v \Phi_l)^{-1} B_X + 2(\Phi_l^* R_v \Phi_l)^{-1} \Lambda_l A_X. \tag{51}$$

Igualando a cero la ecuación anterior

$$0 = -2(\Phi_l^* R_v \Phi_l)^{-1} B_X + 2(\Phi_l^* R_v \Phi_l)^{-1} \Lambda_l A_X, \tag{52}$$

despejando Λ

$$\hat{\Lambda}_l = B_X A_X^{-1}. \tag{53}$$

Se obtiene la siguiente estimación para $\Phi_l^* R_v \Phi_l$ a partir de (46) y (53)

$$\widehat{\Phi_l^* R_v \Phi_l} = \frac{C_X - B_X A_X^{-1} B_X^T}{T}. \tag{54}$$

Medición de errores de estimación y de predicción

La calidad de la estimación obtenida en términos del error cuadrático medio funcional (ECMF) se calcula como

$$ECMF = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} (X_t(i,j) - \hat{X}_t(i,j))^2. \tag{55}$$

para t el tiempo y N número de localizaciones espaciales.

Los errores cuadráticos funcionales obtenidos en la medición de calidad de estimación de los parámetros Q y Λ en la predicción, está dado por la norma Hilbert-Schmidt de las matrices Q y Λ diferenciadas.

$$\begin{aligned}\|Q-\hat{Q}\| &= \frac{1}{l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}\Big(Q(i,j)-\hat{Q}(i,j)\Big)^2,\\ \|\Lambda-\hat{\Lambda}\| &= \frac{1}{l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}\Big(\Lambda(i,j)-\hat{\Lambda}(i,j)\Big)^2,\end{aligned} \tag{56}$$

donde $\hat{\Lambda}$ y $\hat{Q}$ son los valores de Λ y Q estimados respectivamente para alguna iteración del algoritmo EM. Se considera una buena estimación al obtener una norma $< 10^{-4}$.

Implementación computacional

En el presente trabajo se utilizan los lenguajes de programación R 4.3 (R Core Team [2020]) utilizando el IDE Rstudio (RStudio Team [2020]) y Python 3.9 (Van Rossum and Drake Jr [1995]) utilizando el IDE Visual Studio Code en el sistema operativo Ubuntu 20.04 (Sobell [2015]).

R

En el lenguaje R en su versión 4.3 se realiza todo el ETL del conjunto de datos y se utiliza el paquete MissForest (Stekhoven [2022]) para llevar a cabo la imputación de los datos faltantes en cada estación. Adicionalmente se realizan la estadística descriptiva y exploratoria por medio de gráficos de visualización de datos y geográficos con la librería ggplot2 (Wickham [2016]) además del análisis de resultados.

Python

En el lenguaje Python en su versión 3.9 se realizan los cálculos del $SARH(1)$ con ayuda de la biblioteca numpy (Harris et al. [2020]) ya que maneja arreglos multilineales como los tensores y la biblioteca pandas (McKinney et al. [2010]) para la importación del conjunto de datos ya imputado.

Ubuntu 20.04

El sistema operativo utilizado es una distribución de Linux basado en Debian llamado Ubuntu en su versión 20.04 (Sobell [2015]). Este sistema operativo es open-source y es uno de las distribuciones de Linux que tiene más colaboración para el trabajo con datos. La máquina que se utilizó cuenta con:

- 24 Gigas de RAM.
- Procesador Intel® Core™ i5-3450 CPU @ 3.10GHz × 4.

Resultados

En todas las estaciones hubo vacíos de información (337176 datos faltantes en total). Después de imputadas las series, se tienen 48168 registros por estación. Un análisis global de localización y variabilidad de las series es mostrado en las figuras 5 y 6. Aquí se aprecia que en varias de las estaciones existen concentraciones atípicas. Además, la media y la variabilidad cambian espacialmente. Se puede apreciar que las concentraciones de $PM_{2,5}$ más altas se registran en Bosa, Carvajal y Kennedy. Así mismo, se puede concluir que estas tres estaciones, junto con Usme, registran las dispersiones más altas en contraposición a Usaquén, Suba y San Cristóbal, que registran las dispersiones más bajas.

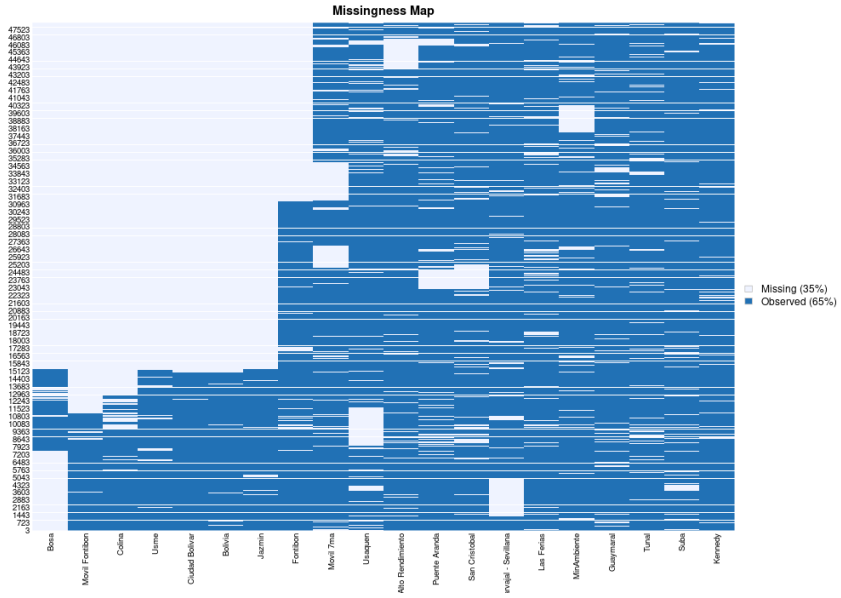

Figura 5: Gráfico de valores faltantes de $PM_{2,5}$ en las estaciones meteorológicas en la Sabana de Bogotá (2017-2020).

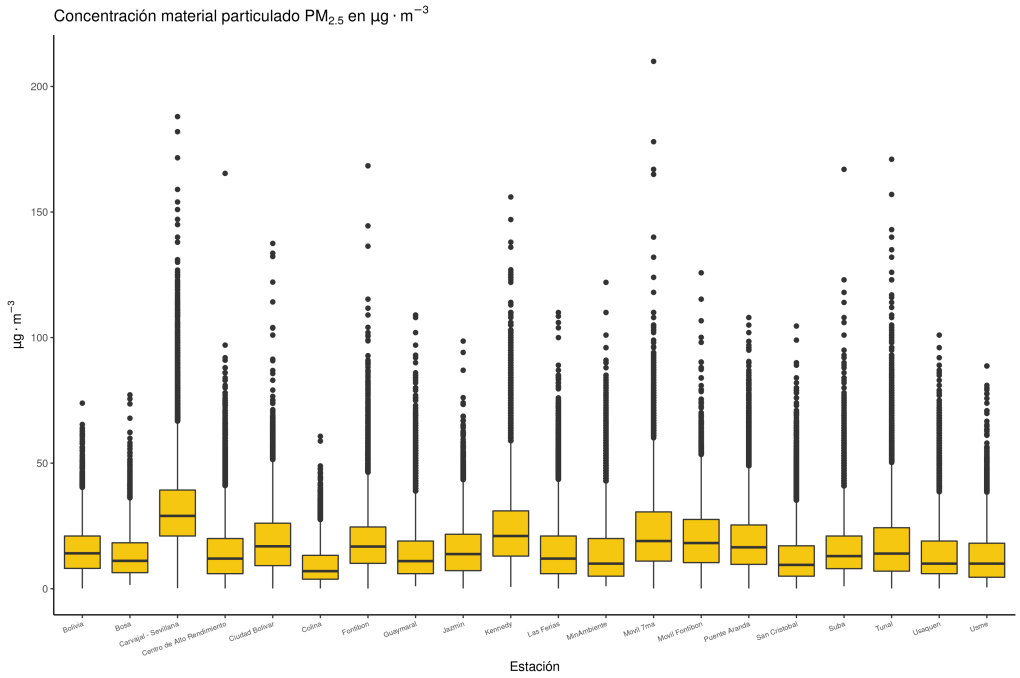

Figura 6: Gráfico de caja de las concentraciones de $PM_{2,5}$ en las estaciones meteorológicas en la Sabana de Bogotá (2017-2020).

En concordancia, al realizar la imputación de los datos, podemos comparar la distribución de la concentración de $PM_{2,5}$ por cada estación y con datos faltantes mostrada en la figura 6 con la distribución generada después de la imputación mostrada en la figura 7. Re puede observar que no existe ninguna diferencia aparente y que todos los datos imputados toman valores de concentración mayores a 0 y sin salirse del dominio de su misma distribución.

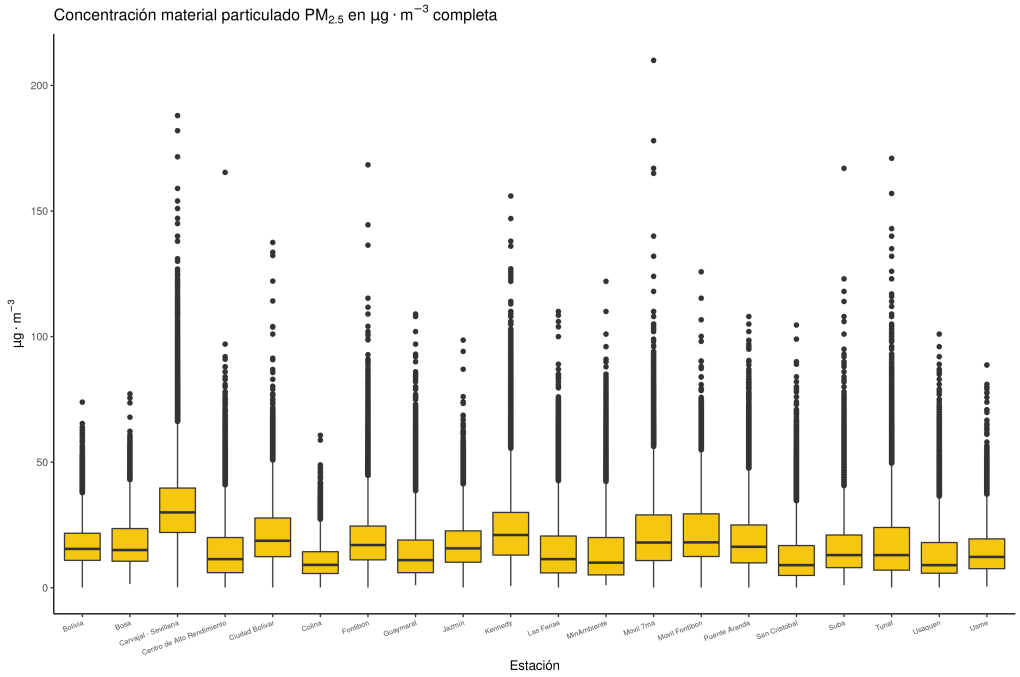

Figura 7: Gráfico de caja de las concentraciones de $PM_{2,5}$ en las estaciones meteorológicas en la Sabana de Bogotá (2017-2020) con datos ya imputados.

De igual manera, comparando las series temporales de cada estación antes (figura 8) y después (figura 9) de la imputación de datos, se puede observar que las series se mantienen a pesar de su imputación y se completan de manera adecuada cumpliendo con su tendencia y periodicidad.

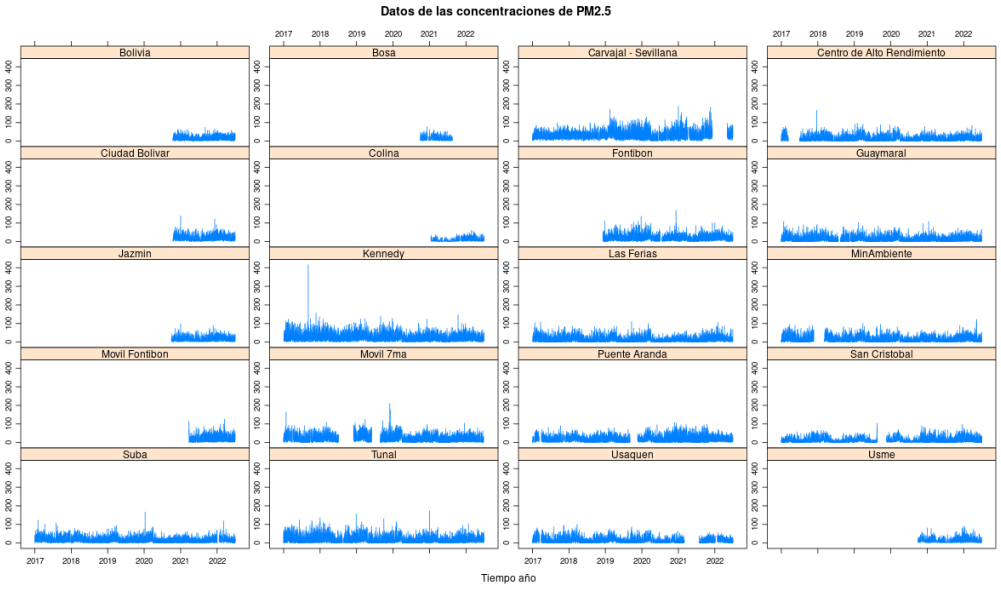

Figura 8: Series temporales de las concentraciones de $PM_{2,5}$ en las estaciones meteorológicas en la Sabana de Bogotá (2017-2020).

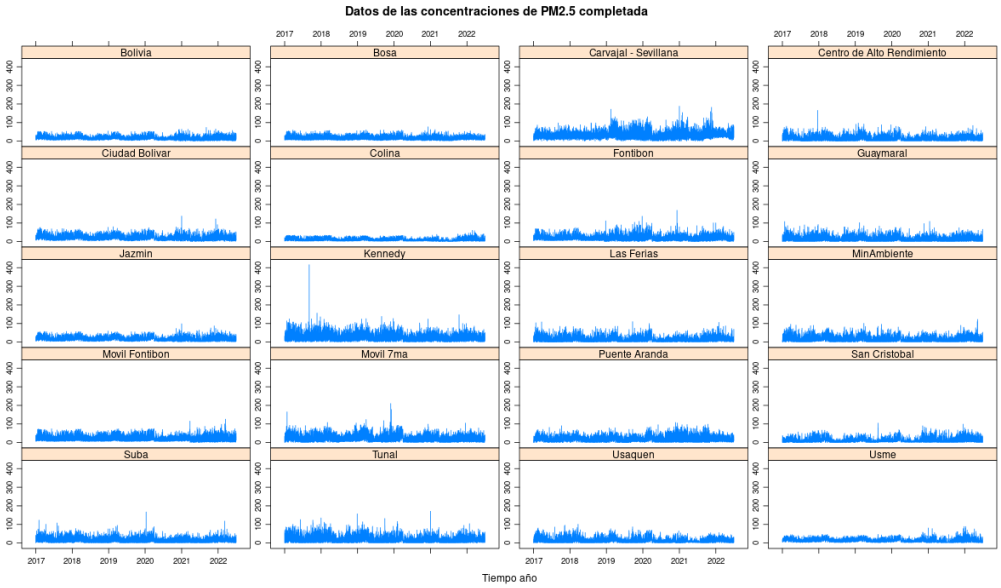

Figura 9: Series temporales de las concentraciones de $PM_{2,5}$ en las estaciones meteorológicas en la Sabana de Bogotá (2017-2020) con datos ya imputados.

Modelo $SARH(1)$

La estimación obtenida de Ψ se presenta en la figura 10, en la que se observa que el operador alcanza su valor máximo en las estaciones de Usme, Carvajal y valor mínimo en Usaquén, Fontibón, Sagrado Corazón y el Tunal; siendo esta última la que registra menores valores en promedio de $PM_{2,5}$ como se evidencia en la figura 6. Por otro lado, las estaciones de Usaquén, Suba y Bolivia presentan mayor homogeneidad en sus concentraciones, lo cual se ve reflejado en el operador estimado Ψ.

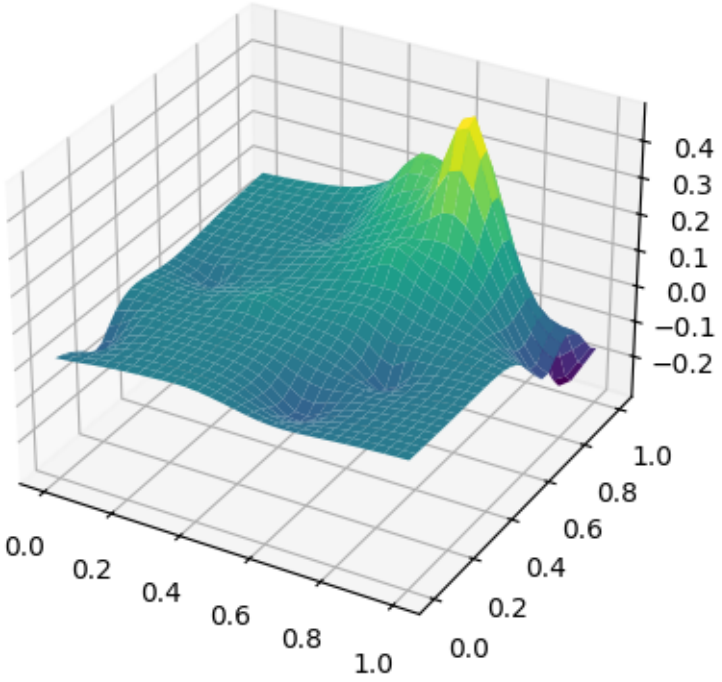

Figura 10: Superficie de la estimación puntual del operador espacio-temporal $\hat{\Psi}$.

Residuales $ECMF$

Con base en $\hat{\Psi}$ se calculan los residuales para $t = 1, ..., 48168$ en las localizaciones muestrales $N = 20$. Se puede medir la calidad de la estimación obtenida en términos del $ECMF$ si presenta una distribución como se muestra en la figura 11, en donde los valores del $ECMF$ se encuentran más concentrados en el orden de 10^{-2}. La estructura funcional y más en su forma autorregresiva, reduce los errores de predicción en su misma esencia L. and P. [2012].

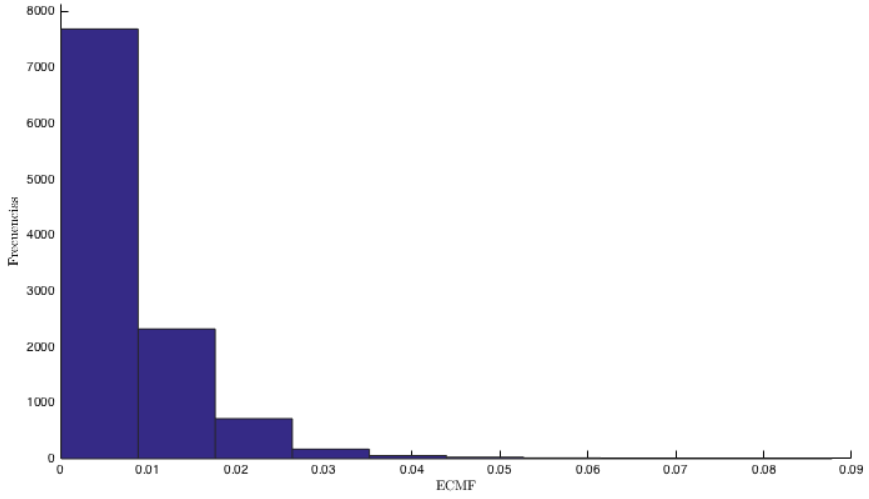

Figura 11: Frecuencias de los $ECMF$ de $\hat{\Psi}$ para $t = 1, ..., 48168$ en las 20 estaciones meteorológicas ubicadas en la sabana de Bogotá.

Los $ECMF$ calculados para $t = 1, ..., 48168$ en las 20 estaciones meteorológicas que se presentan en la figura 11, están comprendidos entre los orden de 1×10^{-5} y 9×10^{-2}.

Validación cruzada

Utilizando el Ψ estimado en el modelo, se hace estimación de las superficies de concentración de $PM_{2,5}$ en los periodos: 11 de enero de 2019 10:00h, 14 de marzo de 2019 10:00h y 2 de agosto de 2019 a las 10:00h. Se calculan los $ECMF$ obtenidos para cada uno de los periodos (tabla 1) enfrentando los valores observados con los predichos.

Cuadro 1: Validación cruzada por estación meteorológica en distintos periodos observados y $ECMF$

	11/01/2019 10:00 horas		14/03/2019 10:00 horas		2/08/2019 10:00 horas	
Estación	**Observado**	**Estimación**	**Observado**	**Estimación**	**Observado**	**Estimación**
Guaymaral	34.3	28.9	35	31.9	39	35.5
Usaquén	40	46	23	20	15	13.1
Suba	30	24	70	81.3	15	15.1
CDAR	53.1	51.2	31.5	34.2	15	13.2
Las Ferias	52	56.4	51.2	45.1	16.7	14.1
Fontibón	71.4	75.8	67.5	49.4	28.6	18.7
Min. Ambiente	73.4	68.4	53.2	61.1	98.9	100.4
Séptima	16	14.1	13	13.1	14	14.7
Puente Ara	31	31.1	34	34.2	28	27.2
Kennedy	12	11	1	12.1	25	26.1
Carvajal	60	67	74	72	47	44
Tunal	40	43	50	43	51	46
San Cristó	22	21	31	29.9	48	46.6
ECMF	$1{,}5 \times 10^{-2}$		$2{,}5 \times 10^{-2}$		$3{,}5 \times 10^{-2}$	

Además, se presenta una gráfica de superficie residual para el 2 de agosto de 2019 a las 10:00h (figura 12), mostrando las diferencias cuadráticas entre los valores observados y los predichos por el modelo, cumpliendo a simple vista con el supuesto de esperanza 0, aleatoridad e independencia.

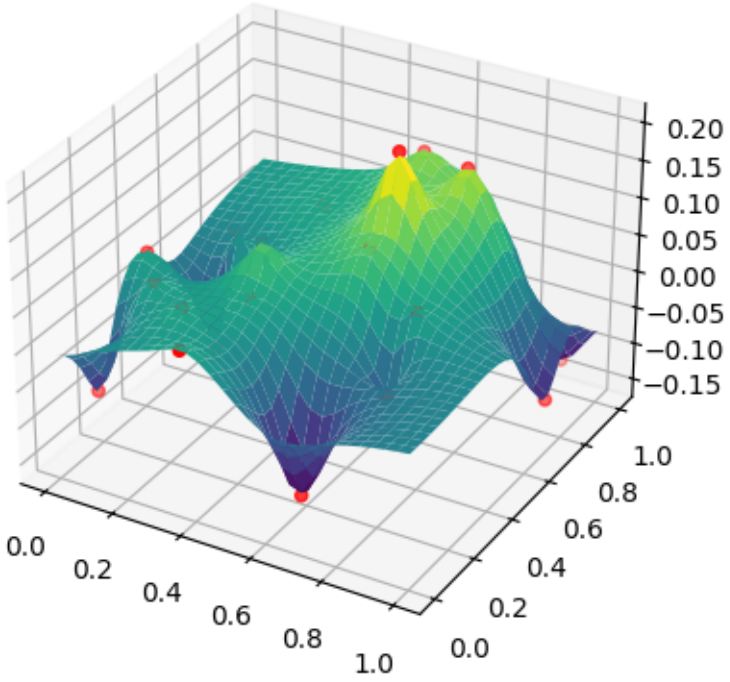

Figura 12: Superficie residual puntual de la concentración de $PM_{2,5}$ para el 2 de agosto de 2019 a las 10:00h en la Sabana de Bogotá.

Conclusiones y Trabajos Futuros

Discusión

Este trabajo indicó que este modelo predictivo permite tener una precisión bastante adecuada de los valores futuros que puede tomar, en este caso, la concentración de $PM_{2,5}$ en el aire respirable del área metropolitana de Bogotá, Colombia, en el corto plazo, con ventanas de tiempo en intervalos de 1 hora y locaciones en donde no existen estaciones meteorológicas. Este modelo permite también predecir eventos de alta concentración en regiones específicas con el fin de tomar decisiones y, reaccionar rápida y eficientemente ante los mismos. La principal contribución de este trabajo radica en el hecho de que utiliza una técnica de ML para imputar datos espacio temporales faltantes para, por medio del $SARH(1)$, señalar la concentración estimada de una variable, que se comporta como una sustancia difundiéndose en un fluido, en una ubicación precisa.

El hecho de ser un modelo autorregresivo hace que la facilidad de la implementación lo convierta en eficiente y de gran utilidad en el campo de los modelos y técnicas para datos espacio temporales funcionales.

Los modelos de series temporales Díaz and Marrón [2013] se desarrollan en una única ubicación espacial con una variable evolucionando en el tiempo y se usan bastante en el sector económico y financiero. Por otro lado, los modelos espaciales Zavala et al. [2006] se desarrollan en un único tiempo con una o varias variables interactuando en diferentes ubicaciones espaciales, siendo estos utilizados en campos como la biología y ecología. También existen los modelos geoestadísticos Axis-Arroyo et al. [2003] que se aproximan bastante a los modelos espacio temporales. Estos trabajan por individual las componentes temporal y espacial de los datos, reflejando esa necesidad de las interacciones espaciales y temporales simultaneas del comportamiento de una variable. En física, existe un comportamiento bien estudiado llamado difusión Mantell et al.. Las sustancias tienden a difundirse en un fluido. Para matemáticamente realizar el modelado determinístico de la difusión, se necesitan condiciones iniciales temporales, espaciales y de características de la sustancia y el fluido. El problema con este modelo determinístico de la difusión es que se convierte en una ecuación diferencial parcial que debe resolverse por métodos numéricos Millán et al. [2011] para cada unidad temporal y espacial, lo que hace que sea demasiado costoso computacionalmente ya que, en la mayoría de las veces, no es posible encontrar una solución analítica. Todas estas metodologías llevan a la necesidad de desarrollar el modelo propuesto en este trabajo, para variables que se comportan como una sustancia difundiéndose en un fluido, que tiene en cuenta la espacio temporalidad simultáneamente y reduce bastante las limitaciones de la solución de la ecuación de difusión. Por esto es una propuesta de gran relevancia sobre los modelos de series temporales, geoestadísticos y físicos, ya que la implementación es sencilla, aumenta la resolución de los datos y consume poco recurso computacional como ya se vió con los resultados obtenidos en este trabajo. El aporte al estudio de problemáticas ambientales, financieras y económicas principalmente, que cumplen los supuestos manifestados anteriormente del modelo predictivo autorregresivo, se considera aceptable a raíz de la validación del mismo que cumple las expectativas en los modelos estadísticos.

En trabajos presentados por Xing et al. [2016], García et al. [2006] y Sanhueza et al. [2006], se evidencia, indistintamente de la ubicación espacial, el impacto que tiene la contaminación del aire respirable, por material

particulado, frente a efectos adversos en la salud y la mortalidad de las personas. Otros autores como León [2009], Barrera [2019] y Grajales [2002], proponen modelos estocásticos y de física estadística complejos para estudiar el comportamiento difusivo en economía financiera. Así mismo, Zuluaga Gómez et al. [2021] y Mesa Mazo et al. [2010], plantean procesos difusivos en fenómenos biológicos y González et al. [2009], soluciones por elementos finitos a modelos difusivos biológicos. De acuerdo a estos antecedentes, resulta propicio emplear, como herramienta de pronóstico y solución, el modelo propuesto en este trabajo, pues es de fácil implementación. Los alcances en temas difusivos del modelo son diversos. No obstante, en trabajos futuros se pretende complementar el modelo mejorando su eficiencia con la integración de tópicos asociados a escenarios difusivos complejos.

Conclusiones

La implementación de los modelos de predicción, de material particulado en el aire, hoy en día, no es más que un vivo reflejo de la búsqueda de instrumentos cada vez más eficientes en la predicción de variables espacio temporales con características de difusión correspondientes a las necesidades que surgen a raíz de la temática de cambio climático y calentamiento global. El uso de energías alternativas y limpias se encuentra en auge, convirtiéndose en la cátedra más común en escuelas, universidades y, por su puesto, en los más altos estamentos gubernamentales.

Como conclusiones se tienen:

- Se logra establecer un modelo Espacio temporal autorregresivo en el espacio de Hilbert que permite predecir la concentración del contaminante $PM_{2,5}$ en el aire del área metropolitana de Bogotá.
- La combinación de técnicas de Machine Learning y modelos estadísticos espacio temporales potencia en gran medida los resultados en el campo de la predicción. En este caso, se evidencia que la técnica Missforest aumenta la resolución de los datos por medio de la imputación, generando disminución en los $ECMF$ llegando al orden medio de 10^{-3} en el $SARH(1)$.
- Es posible realizar una estimación de la concentración de $PM_{2,5}$ en otros puntos del espacio, es decir, en localidades aledañas, a partir del modelo $SARH(1)$ y de la curva de estimación $\hat{X}_t(s)$.
- Este modelo, a diferencia del de series de tiempo o el geoestadístico, involucra todas las observaciones disponibles.
- Por la naturaleza del problema y la postulación del modelo, conjunto con la técnica de imputación, permite implementarse en demás problemáticas tangibles y no tangibles que cumplan los supuestos del modelo siendo su comportamiento principal la difusión homóloga a la de la difusión de una sustancia en un fluido, como por ejemplo, los demás contaminantes presentes en el área metropolitana de Bogotá.

Bibliografía

J Axis-Arroyo, J Mateu, and D Torruco. Diferencias entre modelos geoestadísticos aplicados en el análisis de la distribución espacio-temporal de especies biológicas. 2003.

José Rafael Caro Barrera. Aplicaciones de la física estadística en la valoración de activos financieros: De la ecuación de fokker-planck al modelo de black-scholes. solución en diferencias finitas para una opción put europea. *Estudios de economía aplicada*, 37(2):6–21, 2019.

D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Springer Science, New York, USA, 149 edition, 2012.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. ISSN 1573-0565. doi: https://doi.org/10.1023/A:1010933404324. URL `https://link.springer.com/article/10.1023/A:1010933404324`.

Montserrat Díaz and María del Mar Llorente Marrón. *Econometría*. Ediciones Pirámide, 2013.

Shahla Faisal and Gerhard Tutz. Nearest neighbor imputation for categorical data by weighting of attributes. *Information Sciences*, 592:306–319, 2022. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2022.01.056. URL `https://www.sciencedirect.com/science/article/pii/S0020025522000895`.

Vieu P. Ferraty, F. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York, USA, 149 edition, 2006.

Francisco Fernando García, Rubén Alberto Agudelo, and Karen Margarita Jiménez. Distribución espacial y temporal de la concentración de material particulado en santa marta, colombia. *Revista Facultad Nacional de Salud Pública*, 24(2):73–82, 2006.

Libardo Andrés González, Juan Carlos Vanegas, and Diego Alexander Garzón. Solución numérica de modelos biológicos de reacción difusión en dominios fijos mediante el método de los elementos finitos. *Revista Facultad de Ingeniería Universidad de Antioquia*, (48):65–75, 2009.

Viviana Vélez Grajales. *Difusión y saltos en un modelo de equilibrio general*. El Colegio de México, 2002.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

Meteorología y Estudios Ambientales IDEAM Instituto de Hidrología. Hoja metodológica del indicador Índice de calidad del aire. 1:8, 2012. URL `http://www.ideam.gov.co/documents/24155/125494/35-HM+%C3%8Dndice+calidad+aire+3+FI.pdf/6c0c641a-0c9a-430d-9c37-93d3069c595b`.

Joanna Kaminska. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in wrocław. *Journal of environmental management*, 217:164–174, 03 2018. doi: 10.1016/j.jenvman.2018.03.094.

Horvath L. and Kokoszka P. Inference for functional data with applications. *Springer Science & Business Media*, 2012.

Carlos León. Una aproximación teórica a la superficie de volatilidad en el mercado colombiano a través del modelo de difusión con saltos. *Borradores de Economía; No. 570*, 2009.

C Mantell, M Rodríguez, and E Martínez de la Ossa. coeficientes de difusión por cromatografía supercrítica.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. *Advances in neural information processing systems*, 12, 1999.

Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

Mónica Jhoana Mesa Mazo, Paulo César Tintinago Ruiz, Carmen Alicia Ramírez Bernate, Alejandra María Pulgarín Galvis, Omar Alejandro Arse Serna, and Oscar Emilio Molina Díaz. *Efectos de la difusión de un contaminante en la dinámica y la dispersión poblacionales en un medio acuático: Modelado y aproximación.*. PhD thesis, Ciencias Básicas y Tecnologías-Maestría en Biomatemáticas, 2010.

Zulma Millán, Leonor de la Torre, Laura Oliva, and María Del Carmen Berenguer. Simulación numérica: Ecuación de difusión. 2011.

Shigeyuki Oba, Masa-aki Sato, Ichiro Takemasa, Morito Monden, Ken-ichi Matsubara, and Shin Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 11 2003. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg287. URL `https://doi.org/10.1093/bioinformatics/btg287`.

OMS Organización Mundial de la Salud. Directrices mundiales de la oms sobre la calidad del aire: partículas en suspensión ($pm_{2,5}$ y pm_{10}), ozono, dióxido de nitrógeno, dióxido de azufre y monóxido de carbono. 2021. URL `https://iris.who.int/bitstream/handle/10665/346062/9789240035461-spa.pdf?sequence=1&isAllowed=y`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

J. Ramsay. *Functional Data Analysis*. John Wiley & Sons, NY, USA, Second Edition edition, 2006.

Nagy S. Riesz, F. Functional analysis. *New York*, 3(6):35, 1990.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020. URL `http://www.rstudio.com/`.

Salmerón R. Ruiz, M. and J. Angulo. Kalman filtering from pop-based diagonalization of arh (1). *Computational Statistics & Data Analysis*, 51(10):4994–5008, 2007.

M. D. Ruiz-Medina. Spatial functional prediction from spatial autoregressive hilbertian processes. *Environmetrics*, 23(1):119–128, 2012.

María Dolores Ruiz-Medina. Spatial autoregressive and moving average hilbertian processes. *Journal of Multivariate Analysis*, 102(2):292–305, 2011.

MD Ruiz-Medina and RM Espejo. Spatial autoregressive functional plug-in prediction of ocean surface temperature. *Stochastic environmental research and risk assessment*, 26(3):335–344, 2012.

Pedro Sanhueza, Claudio Vargas, and Paula Mellado. Impacto de la contaminación del aire por pm10 sobre la mortalidad diaria en temuco. *Revista médica de Chile*, 134(6):754–761, 2006.

SDAB Secretaría Distrital de Ambiente de Bogotá. Plan estratégico para la gestión integral de la calidad del aire de bogotá 2030. 1, 2021. URL `https://drive.google.com/file/d/1Pt7cGCRSzm8ogsA450Tauy0J065ZU9nW/view`.

Mark G Sobell. *A practical guide to Ubuntu Linux*. Pearson Education, 2015.

Daniel J. Stekhoven. *missForest: Nonparametric Missing Value Imputation using Random Forest*, 2022. R package version 1.5.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr597. URL `https://doi.org/10.1093/bioinformatics/btr597`.

Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Yu-Fei Xing, Yue-Hua Xu, Min-Hua Shi, and Yi-Xin Lian. The impact of pm2. 5 on the human respiratory system. *Journal of thoracic disease*, 8(1):E69, 2016.

Miguel A Zavala, Rubén Díaz-Sierra, Drew Purves, Gustazo E Zea, and Itziar R Urbieta. Modelos espacialmente explícitos. *Ecosistemas*, 15(3), 2006.

Oscar Arley Zuluaga Gómez, Jorge Eduardo Patiño Quinchía, and German Mauricio Valencia Hernández. Modelos implementados en el análisis de series de tiempo de temperatura superficial e índices de vegetación: una propuesta taxonómica en el contexto de cambio climático global. *Revista de Geografía Norte Grande*, (78):323–344, 2021.