Universidad de **Valladolid**

Facultad de Ciencias

Trabajo fin de grado

Grado en Física

Predicción de concentraciones de contaminantes atmosféricos mediante modelos estadísticos

Autor: Javier Luna Manteca

Tutores: Isidro A. Pérez Bartolomé, Mª Ángeles García Pérez

Octubre 2023

Índice

1. Introducción **4**

2. Materiales y métodos **5**
2.1. Datos de concentración de contaminantes atmosféricos 5
2.2. Justificación de la clasificación en regiones 6
2.3. Análisis estadístico en R 7
2.4. Modelo de predicción con Prophet 8
2.4.1. Breve introducción a Prophet 8
2.4.2. Primer acercamiento 10
2.4.3. Segundo acercamiento 10
2.4.4. Regresores 11
2.5. Evaluación de los resultados 11

3. Resultados **11**
3.1. Análisis estadístico y predicción del dióxido de nitrógeno o NO_2 11
3.1.1. Predicción con el algoritmo Prophet. Primer acercamiento 17
3.1.2. Predicción con el algoritmo Prophet. Segundo acercamiento 22
3.2. Análisis estadístico y predicción del ozono u O_3 24
3.2.1. Predicción con el algoritmo Prophet. Segundo acercamiento 28
3.3. Análisis estadístico y predicción de las PM_{10} 29
3.3.1. Predicción con el algoritmo Prophet. Segundo acercamiento 32
3.4. Análisis estadístico y predicción de las $PM_{2,5}$ 34
3.4.1. Predicción con el algoritmo Prophet. Segundo acercamiento 36

4. Conclusiones **38**

5. Referencias **42**

Anexo A **I**

Anexo B **XIII**

Anexo C **XXV**

Anexo D **XXXVII**

Abstract

Este trabajo aborda el análisis estadístico de series temporales de la concentración de contaminantes atmosféricos en tres regiones diferentes del núcleo urbano de Londres. En particular se utilizan datos del dióxido de carbono (NO_2), el ozono (O_3) y las partículas en suspensión PM_{10} y $PM_{2,5}$ entre los años 2007 y 2011. Se emplea el algoritmo Prophet de Facebook en R para obtener modelos estadísticos ajustados a los datos proporcionados y realizar predicciones basadas en estos. Se observa una fuerte correlación entre las regiones y se identifican patrones estacionales en los contaminantes. El ozono muestra comportamientos inversos. Se evalúan dos acercamientos con el algoritmo Prophet, destacando la importancia del número de días utilizados para ajustar el modelo. Las conclusiones incluyen tendencias estacionales, ciclos anuales y semanales, y el poco número de parámetros que se necesitan para obtener predicciones precisas.

This study addresses the statistical analysis of time series data on the concentration of atmospheric pollutants in three different regions within the urban core of London. Specifically, data on nitrogen dioxide (NO_2), ozone (O_3), and particulate matter (PM_{10} and $PM_{2,5}$) from the years 2007 to 2011 are utilized. The Facebook Prophet algorithm in R is employed to develop statistically adjusted models based on the provided data and to make predictions accordingly. A strong correlation is observed among the regions, and seasonal patterns in pollutants are identified. Ozone exhibits inverse behaviors. Two approaches are evaluated with the Prophet algorithm, emphasizing the significance of the number of days used for model adjustment. Conclusions include seasonal trends, annual and weekly cycles, and the low number of parameters needed for achieving accurate predictions.

1. Introducción

Dentro de los contaminantes cuyo impacto se percibe en el medio urbano, el dióxido de nitrógeno (NO_2), el ozono (O_3) y las partículas o materia particulada (PM) ocupan un lugar destacado.

El NO_2 se origina en los procesos de combustión, como es el caso de los vehículos y también en instalaciones industriales, como las de generación eléctrica. Se sabe que, a escala global, entre 2000 y 2019, las concentraciones de NO_2 crecieron en el 71 % de las áreas urbanas (Sicard *et al.*, 2023).

El O_3 carece de fuentes naturales en la troposfera, por lo que su presencia se debe a distintos mecanismos. El primero está formado por reacciones fotoquímicas de gases precursores, como son los compuestos orgánicos volátiles y el monóxido de carbono en presencia de óxidos de nitrógeno (NO y NO_2, denominados NO_x en conjunto). Estas reacciones son no lineales y se ven favorecidas cuando las temperaturas son altas (Liu *et al.*, 2023). El segundo mecanismo responsable de las altas concentraciones de O_3 son las intrusiones estratosféricas. En este caso el O_3 se propaga hacia abajo en la troposfera hasta, en algunos casos, alcanzar la superficie (Roy *et al.*, 2023). Por último, también se observa O_3 por el transporte de concentraciones durante la noche en una capa residual situada entre 200 y 500 m sobre el suelo (Guo *et al.*, 2024). Estas concentraciones pueden alcanzar la superficie al día siguiente cuando se desarrolle una capa límite convectiva.

En cuanto a las partículas o materia particulada, típicamente se distingue entre partículas de diámetro inferior a 10 μm, que se denotan por PM_{10} , y partículas de tamaño inferior a 2,5μm, que se denotan como $PM_{2,5}$. La materia particulada es una mezcla dispersa en la atmósfera que puede incluir partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen. Su origen puede deberse a procesos naturales, como el volcán de La Palma (Milford *et al.*, 2023), o el polvo del desierto (Birinci *et al.*, 2023), pero también se originan en actividades realizadas por el hombre como el transporte o la minería (Kholodov *et al.*, 2022; Palei *et al.*, 2023).

Este trabajo considera la distribución espacial de estos cuatro contaminantes en un núcleo urbano, la ciudad de Londres. En ella se han establecido tres áreas de acuerdo con la cuantía de las concentraciones. Con este punto de partida, el trabajo tiene como objetivo analizar estadísticamente las concentraciones de estas tres zonas para ver sus contrastes, pero, además, considera la predicción de dichas concentraciones.

Para llevar a cabo esta predicción, se ha empleado una herramienta muy útil en el análisis de series temporales, el algoritmo Prophet (Taylor *et al.*, 2021), desarrollado por Facebook. Prophet se destaca por su capacidad para manejar datos de series temporales con tendencias, estacionalidad y días festivos, características comunes en el análisis de la calidad del aire urbano. Es capaz de adaptarse tanto a datos faltantes en la serie como a datos atípicos.

Para la ejecución de este análisis, predicción y la exposición de los resultados de manera gráfica se ha utilizado el lenguaje de programación R (R Core Team, 2023), ampliamente conocido en la comunidad estadística.

2. Materiales y métodos

En esta sección se describirán los materiales utilizados, así como los métodos empleados para llevar a cabo el análisis estadístico de los datos de concentración de núcleos contaminantes en la atmósfera de Londres y la posterior predicción de valores utilizando el algoritmo Prophet de Facebook en el entorno de programación R.

2.1. Datos de concentración de contaminantes atmosféricos

Para este estudio se han utilizado cuatro archivos de datos con extensión .DAT que contienen información sobre la concentración de los contaminantes atmosféricos NO_2, O_3, PM_{10} y $PM_{2,5}$ en Londres.

Estos datos corresponden a las 7681 parejas de latitudes y longitudes comprendidas entre las coordenadas 51,08 N y 51,83 N y 0,83 W y 0,50 E. Estos puntos se pueden observar en la Figura 1.

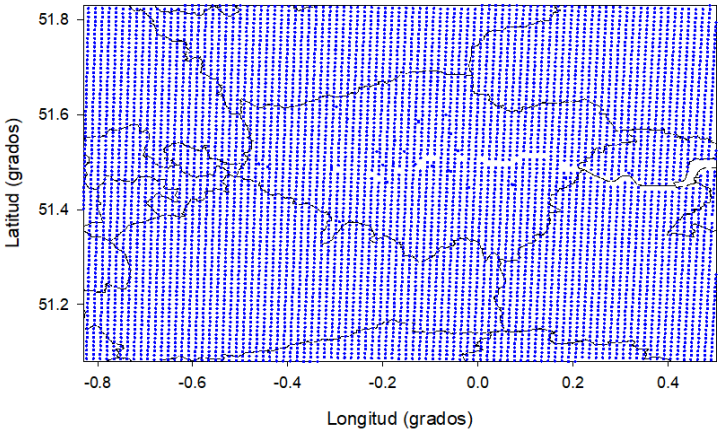


Figura 1: Puntos utilizados para calcular los datos utilizados en el estudio.

Cada archivo de datos está estructurado en cinco columnas, que incluyen el año, el mes, el día, la concentración y la región correspondiente a cada dato.

Las regiones de Londres se clasificaron en tres categorías (1, 2 y 3) en función de las concentraciones registradas para cada contaminante. Esta columna de región permitirá la segmentación de los datos y su análisis posterior.

2.2. Justificación de la clasificación en regiones

Las concentraciones se han suavizado mediante la expresión

$$c(x, y, h_1, h_2) = \frac{\sum_{i=1}^{N} \left(\frac{x-x_i}{h_1}\right) K_2 \left(\frac{y-y_i}{h_2}\right) c_i}{\sum_{i=1}^{N} \left(\frac{x-x_i}{h_1}\right) K_2 \left(\frac{y-y_i}{h_2}\right)} \tag{1}$$

donde c es la concentración en el punto (x, y), c_i es la concentración conocida en el punto (x_i, y_i), K_i son los núcleos del suavizado, h_i son las ventanas. Se ha empleado un núcleo gaussiano

$$K(t) = (2\pi)^{-\frac{1}{2}} \exp\left(-0{,}5t^2\right) \; -3 < t < 3, \tag{2}$$

donde el intervalo usado para calcularlo se ha limitado para aumentar la velocidad de cálculo (Fernández-Duque *et al.*, 2019). La ventana h se ha calculado siguiendo el método de Silverman (Donnelly *et al.*, 2011).

$$h = 0{,}9\sigma n^{-\frac{1}{5}} \tag{3}$$

donde σ es la desviación estándar. Se han considerado como concentraciones de fondo los promedios de los extremos NE, SE, SW, NW de la región, y el intervalo de concentración desde el fondo al máximo se ha dividido en tres partes iguales, que son las isolíneas que se representan en la Figura 2.

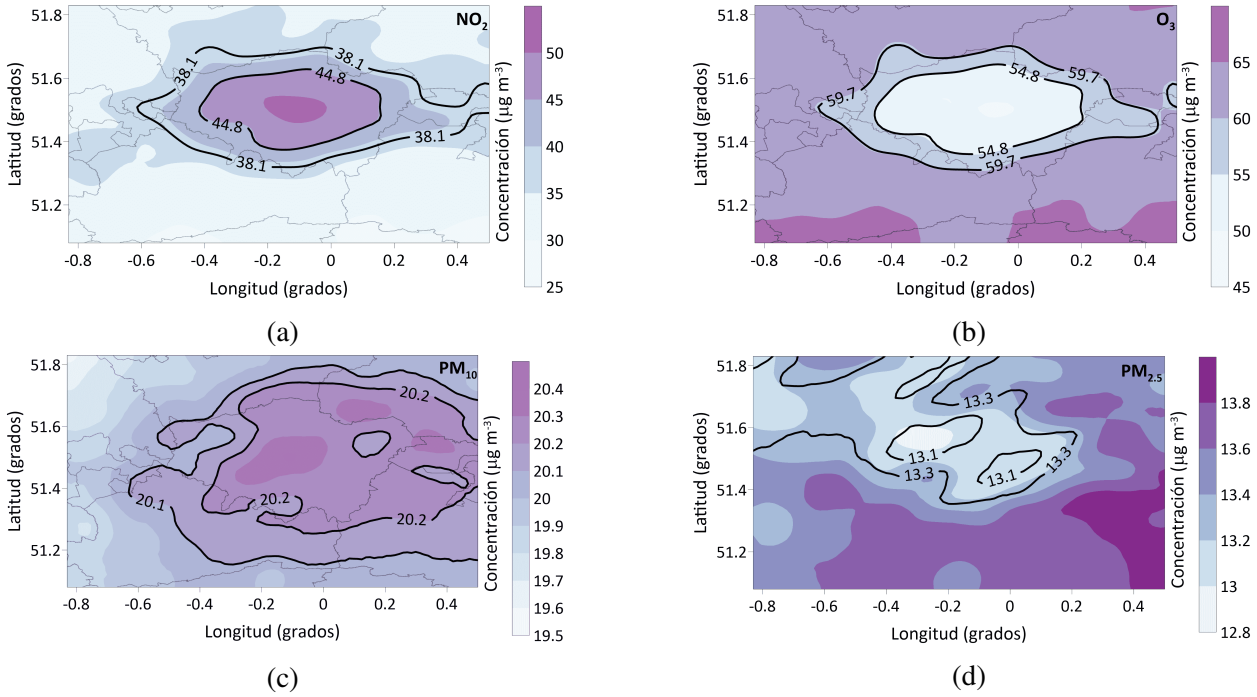


Figura 2: (Regiones calculadas para los datos de concentración de núcleos de NO_2 (a), O_3 (b), PM_{10} (c) y $PM_{2,5}$ (d).

2.3. Análisis estadístico en R

El análisis estadístico de los datos se llevará a cabo utilizando el lenguaje de programación R y las librerías necesarias en el entorno de desarrollo RStudio. Estas librerías son *dplyr* (Wickham *et al.*, 2023), *lubridate* (Grolemund *et al.*, 2011), *timeDate* (Wuertz *et al.*, 2023), *prophet* (Taylor *et al.*, 2021) y *ggplot2* (Wickham, 2016). A continuación se describen los pasos principales del análisis:

- Importación y lectura de datos: Los archivos de datos de concentración de los diferentes contaminantes serán importados en R para su manipulación y análisis. Se utilizarán las funciones adecuadas para cargar los datos en un formato estructurado de dataframe que permita su procesamiento posterior. Se utilizan métodos como *read.table*, *as.data.frame.matrix* y *make_date*. El dataframe de datos inicial tendrá las mismas columnas que los datos proporcionados. A estas columnas añadiremos una columna de tipo **Date** en la que agruparemos año, mes y día con formato yyyy-mm-dd.

- Manipulación de datos: El dataframe inicial será manipulado para obtener diferentes dataframes que permitan la exploración y el análisis de los datos. Se harán diferentes combinaciones de los datos por regiones, por años, por meses, etc. y se calcularán diferentes estadísticos como la media aritmética y la mediana. Para esta manipulación utilizamos en profundidad diferentes métodos de la librería *dplyr*, como son *filter*, *group_by*, *summarise*, *mutate* . . .

- Exploración de datos para obtener información descriptiva: Se utilizarán gráficos y tablas para visualizar y resumir los datos. Primero se representarán los datos de la concentración en función de la fecha con el método *plot*. Después se hará uso del método *boxplot* para representar diagramas de cajas y bigotes y poder hacer un análisis visual de los datos. Los estadísticos que se van a calcular son la media aritmética, la mediana, el rango intercuartílico (RIC) y el índice de Yule-Kendall.

La **media aritmética** es la suma de un conjunto de valores dividida entre el número total de sumandos. La **mediana** es el número intermedio de un grupo de números; es decir, la mitad de los números son superiores a la mediana y la mitad de los números tienen valores menores que la mediana. El **rango intercuartílico** es la diferencia entre el tercer y el primer cuartil de una distribución. $RIC = q_{0,75} - q_{0,25}$.

El **índice de Yule-Kendall** γ_{YK} será mayor que cero en caso de que los datos tengan una tendencia hacia la derecha, al menos el 50 % de ellos, esto quiere decir que la distancia a la mediana será mayor desde el cuartil superior que desde el cuartil inferior. De forma inversa los datos sesgados a la izquierda tienen un índice Yule-Kendall negativo.

$$\gamma_{YK} = \frac{q_{0,25} - 2q_{0,5} + q_{0,75}}{RIC} \tag{4}$$

- Análisis descriptivo de las concentraciones de los contaminantes atmosféricos en función de las regiones definidas. Se observarán tendencias y periodicidades para poder especificar diferentes parámetros en el algoritmo Prophet de predicción de datos. La *tendencia* se refiere a la dirección general y persistente en la que se mueven los datos a largo plazo. La *periodicidad* se refiere a la presencia de patrones repetitivos en una serie temporal. Es decir, ciertos eventos o fluctuaciones en los datos que ocurren de manera regular a lo largo del tiempo.

- Modelo de predicción con Prophet: Utilizando el algoritmo Prophet de Facebook se realizará la predicción de valores futuros de las concentraciones de los contaminantes atmosféricos en las diferentes regiones. Se ajustará un modelo de series temporales con Prophet y se generarán pronósticos para periodos determinados.

2.4. Modelo de predicción con Prophet

Se han seguido los mismos pasos para predecir los valores de cada contaminante. Tras un análisis estadístico de los datos se ha decidido predecir únicamente los valores de la región uno. Se han realizado dos acercamientos diferentes. Se han representado gráficamente todas las predicciones. Todas las gráficas se pueden encontrar en los anexos. Algunas de las gráficas más representativas se incluyen en el trabajo.

2.4.1. Breve introducción a Prophet

Una *serie temporal* es una colección de observaciones numéricas organizadas en un orden natural. Normalmente cada observación está asociada con un instante particular o intervalo de tiempo, y esto es lo que proporciona el orden de la serie temporal.

El algoritmo de predicción 'Prophet' utiliza un modelo estadístico basado en descomponer la serie temporal en tres componentes principales que representan:

- Tendencia
- Estacionalidad
- Festivos y eventos

Para la tendencia utiliza una función no lineal que se ajusta a los cambios no periódicos a largo plazo. Para la estacionalidad utiliza un modelo de Fourier para capturar patrones repetitivos. La estacionalidad puede ser modelada como componente aditivo o como componente multiplicativo. En este último caso, la componente de la estacionalidad se representa como un factor que multiplica la componente de la tendencia. La componente de los festivos y eventos estudia el impacto que tiene un día festivo específico sobre los valores de la serie temporal.

Estas tres componentes se combinan, junto con el error, de manera aditiva (Taylor *et al.*, 2017):

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{5}$$

Donde,

- La función de tendencia $g(t)$ modela cambios no periódicos en el valor medio de la serie temporal.
- La componente estacional $s(t)$ representa cambios periódicos de periodo conocido (p.ej., estacionalidad semanal y/o anual).
- La componente $h(t)$ representa el efecto de los días festivos o eventos que ocurren de manera irregular.
- La componente aleatoria o término del error ϵ_t representa cualquier cambio que no haya sido registrado por el resto de componentes.

La manera más fácil de ajustar el modelo para predecir los datos con el algoritmo 'Prophet' consiste en pasar al método *prophet* un dataframe con dos columnas, una con la variable tiempo y la otra con los datos correspondientes, y almacenar el resultado en una variable nueva. El nombre de esta variable puede ser por ejemplo **model**. Estas columnas deben ser nombradas como 'ds', en el caso de la variable tiempo, e 'y', en el caso de los datos.

Para hacer la predicción, primero se crea un nuevo dataframe, por ejemplo de nombre **future**, con una única columna, la cual debe contener los valores de tiempo sobre los que se quiera hacer la predicción. Una vez hecho esto, se pasa al método *predict* los argumentos **model** y **future** creados anteriormente. Se almacena el resultado en una nueva variable, por ejemplo **forecast**.

Para representar estos resultados gráficamente se pueden utilizar métodos como *plot*(model, forecast) y *prophet_plot_components*(model, forecast).

Más allá del método más simple para hacer predicciones, el algoritmo Prophet permite al analista especificar multitud de parámetros para ajustar las componentes del modelo. Entre estos parámetros se encuentran especificaciones sobre las estacionalidades o fechas de eventos o días festivos específicos que puedan tener influencia en la serie temporal (Taylor *et al.*, 2021).

2.4.2. Primer acercamiento

En este primer acercamiento se han seleccionado un número determinado de días aleatorios entre los días de los datos proporcionados. Estos días se han denominado días 'aleatorizados'. Se han seleccionado por orden 50, 100, 150 y 200 días. Primero se ha hecho la selección entre los días del año 2007 y después se han añadido escalonadamente los años 2008, 2009, 2010 y 2011 para realizar la selección. En total se ha trabajado con cuatro conjuntos de días para cada uno de los cinco conjuntos de años, es decir veinte conjuntos de días aleatorizados. Se han utilizado los valores de los días de cada uno de estos conjuntos para ajustar veinte modelos para cada contaminante y realizar predicciones en tres pasos diferentes. En total sesenta modelos por contaminante.

En el primer paso se ha ajustado el modelo con los días aleatorizados y se han predicho los valores para los mismos días con los que se ha ajustado el modelo y los días directamente consecutivos a estos.

En el segundo paso se ha ajustado el modelo con:

- los valores de los días aleatorizados,
- los valores de los días directamente consecutivos a los días aleatorizados.

Se han predicho los valores para los mismos días con los que se ha ajustado el modelo y los días directamente consecutivos a estos.

Hay que tener en cuenta que en este segundo paso del primer acercamiento es posible que al escoger los días directamente consecutivos de los días aleatorizados haya días repetidos, ya que es posible que entre los días aleatorizados ya haya días consecutivos. En cualquier caso se tiene que

$$\text{días para ajustar el modelo} \leq 2 \times \text{días aleatorizados}$$

En el tercer y último paso de este primer acercamiento se ha seguido un procedimiento similar al utilizado en el segundo paso pero en este caso se han utilizado los valores predichos en el primer paso para hacer el ajuste. Es decir, se ha ajustado el modelo con:

- los valores de los días aleatorizados,
- los valores predichos de los días directamente consecutivos a partir de un modelo ajustado con los valores de los días aleatorizados.

Las consideraciones hechas para el segundo paso son igualmente válidas en este último paso.

2.4.3. Segundo acercamiento

En este segundo acercamiento se ha ajustado el modelo con valores de días consecutivos y se han predicho los valores de todos los días desde el 1 de enero de 2007 hasta el 31 de diciembre de 2011. Se han ajustado tres modelos diferentes con los valores de todos los días desde el 1 de enero de 2007 hasta:

1. el 30 de junio de 2010,
2. el 31 de julio de 2011,
3. el 30 de noviembre de 2011.

2.4.4. Regresores

El algoritmo Prophet permite añadir 'regresores', esto es, datos adicionales, para mejorar la predicción. En particular, permite añadir como regresores otras series temporales con la misma componente de tiempo que los datos utilizados para realizar el ajuste.

Para la predicción de los valores pertenecientes a la región uno de cada uno de los contaminantes no se han añadido los datos de las otras dos regiones de ese mismo contaminante como regresores. Los datos de las tres regiones de un mismo contaminante siguen trayectorias en el tiempo muy similares y se produce un sobreajuste si se añaden dichos valores.

No se han incluido regresores para ninguno de los ajustes del NO_2, el O_3 y las PM_{10} y tampoco para los del primer acercamiento de las $PM_{2,5}$. Para los ajustes del segundo acercamiento de $PM_{2,5}$ se han añadido los datos de los otros tres contaminantes como regresores.

2.5. Evaluación de los resultados

Para evaluar los resultados del análisis estadístico y las predicciones generadas por el algoritmo Prophet, se utilizará el error absoluto medio (EAM).

El algoritmo suaviza la curva de valores de la serie temporal, lo que significa que evita valores atípicos para generar, en general, una serie uniforme en el tiempo. Por esto, hay valores predichos muy diferentes al valor real (atípico en este caso) que hacen que el error cuadrático medio sea grande y poco representativo. Por esto se ha decidido utilizar el error absoluto medio.

Sea n el número de valores de la serie temporal, $y \in \mathbb{K}^n$ los valores de la predicción y $x \in \mathbb{K}^n$ los valores reales, se presenta en la ecuación (6) el cálculo del error absoluto medio de la predicción.

$$EAM = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{6}$$

3. Resultados

Se exponen los resultados del análisis individual de los diferentes contaminantes atmosféricos y las predicciones de cada uno de ellos. Todos los datos correspondientes a la concentración están medidos en $\frac{\mu g}{m^3}$.

3.1. Análisis estadístico y predicción del dióxido de nitrógeno o NO_2

Una vez importados y leídos los datos, se ha obtenido un dataframe con una entrada por cada día entre el primer día de 2007 y el último día de 2011. A este dataframe se le ha añadido una columna que se ha nombrado 'fecha' para que el dataframe final con los datos de entrada sea como el que se muestra en la Figura 3.

	región	año	mes	día	concentración	fecha
1	1	2007	1	1	33.88300	2007-01-01
2	1	2007	1	2	54.29128	2007-01-02
3	1	2007	1	3	47.44779	2007-01-03
4	1	2007	1	4	45.25599	2007-01-04
5	1	2007	1	5	52.00705	2007-01-05
6	1	2007	1	6	59.91958	2007-01-06
7	1	2007	1	7	29.49229	2007-01-07
8	1	2007	1	8	58.20185	2007-01-08
9	1	2007	1	9	28.22475	2007-01-09
10	1	2007	1	10	52.88555	2007-01-10

Figura 3: Dataframe inicial con los datos de la concentración del NO_2 a lo largo del periodo de estudio.

Este dataframe tiene $365 \times 5 + 1 = 1825 + 1 = 1826$ filas por región, ya que los datos incluyen valores de cinco años, entre los que se incluye 2008, que fue bisiesto. Como hay tres regiones, el número total de filas o entradas del dataframe es $1826 \times 3 = 5478$.

Se representan en la Figura 4 los datos de la concentración a lo largo del tiempo con un color diferente para cada región.

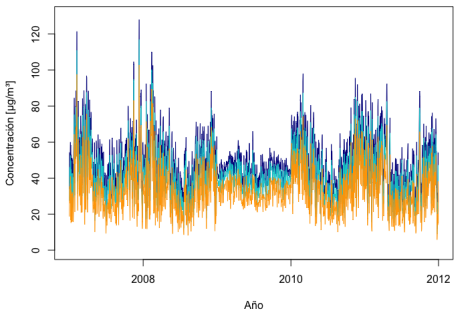


Figura 4: Concentración de NO_2 en la atmósfera del núcleo urbano de Londres en el período de estudio en tres regiones: Región 1 (azul oscuro), Región 2 (azul claro), Región 3 (naranja).

Si se observa la Figura 4, es bastante claro que los datos en las tres regiones siguen la misma tendencia y trayectoria. Los coeficientes de correlación de Pearson se muestran en la Tabla 1.

Tabla 1: Coeficiente de correlación de Pearson calculado entre las regiones combinadas de dos en dos.

Regiones	Pearson
1-2	0,998
1-3	0,984
2-3	0,992

Que los coeficientes de correlación de Pearson sean tan altos permite sacar varias conclusiones, todas con la misma idea o significado:

- Hay una fuerte relación lineal positiva entre los datos de la concentración en las diferentes regiones. Es decir, a medida que los valores de una región varían también varían los valores en las otras regiones de manera consistente y proporcional.
- Los datos de la concentración en las diferentes regiones siguen una trayectoria casi idéntica. Los patrones y las tendencias en ambos conjuntos de datos son muy parecidos.
- Hay una dependencia fuerte entre los valores de las concentraciones en las diferentes regiones. La variación en una región se puede explicar en gran medida por la variación en las otras regiones.

Todo esto lleva a la conclusión de que se pueden analizar los movimientos generales de la serie observando únicamente los valores de una de las regiones, o alternativamente utilizando para cada día el valor medio de las tres regiones.

Se han calculado diferentes estadísticos con estos valores medios en función del año para hacer el análisis descriptivo. En la Tabla 2 se muestran los principales estadísticos de estudio, esto es, medias aritméticas, medianas, rangos intercuartílicos (RIC), e índices de Yule-Kendall.

Tabla 2: Tabla con estadísticos calculados para analizar la evolución general de la serie.

año	media $\left[\frac{\mu g}{m^3}\right]$	mediana $\left[\frac{\mu g}{m^3}\right]$	RIC $\left[\frac{\mu g}{m^3}\right]$	índice Yule - Kendall
2007	43.398	41.125	19.568	0.070
2008	40.722	39.478	21.070	-0.015
2009	38.884	39.393	8.441	-0.103
2010	43.001	42.067	21.127	-0.003
2011	39.513	38.083	22.079	0.033

Los valores de la Tabla 2 permiten llegar a las siguientes conclusiones:

- Que los valores de la media y de la mediana sean similares entre sí indica que hay una distribución equilibrada de valores tanto por encima como por debajo del valor central. Además, se puede decir que los valores están igualmente distribuidos por encima y por debajo de la mediana y que, o no hay gran cantidad de valores atípicos, o están asimismo distribuidos equilibradamente por encima y por debajo de la mediana. En otro caso la media se vería sesgada hacia un extremo. Se puede utilizar tanto la media como la mediana para describir la dirección general que sigue nuestra serie temporal.

- Con los valores de la media se puede ver cómo el valor de la concentración de NO_2 decrece linealmente hasta 2009, en 2010 tiene un pico y decrece en 2011 con mayor rapidez que antes.

Si uno se fija en la Tabla 2, parece que la media está siempre en torno a $40\frac{\mu g}{m^3}$. Se representan en la Figura 5 los datos de las medias diarias de las concentraciones en las tres regiones. En cada gráfica se añade una línea horizontal a la altura de la concentración media correspondiente al año para ver cómo se comporta la serie en torno a ese valor.

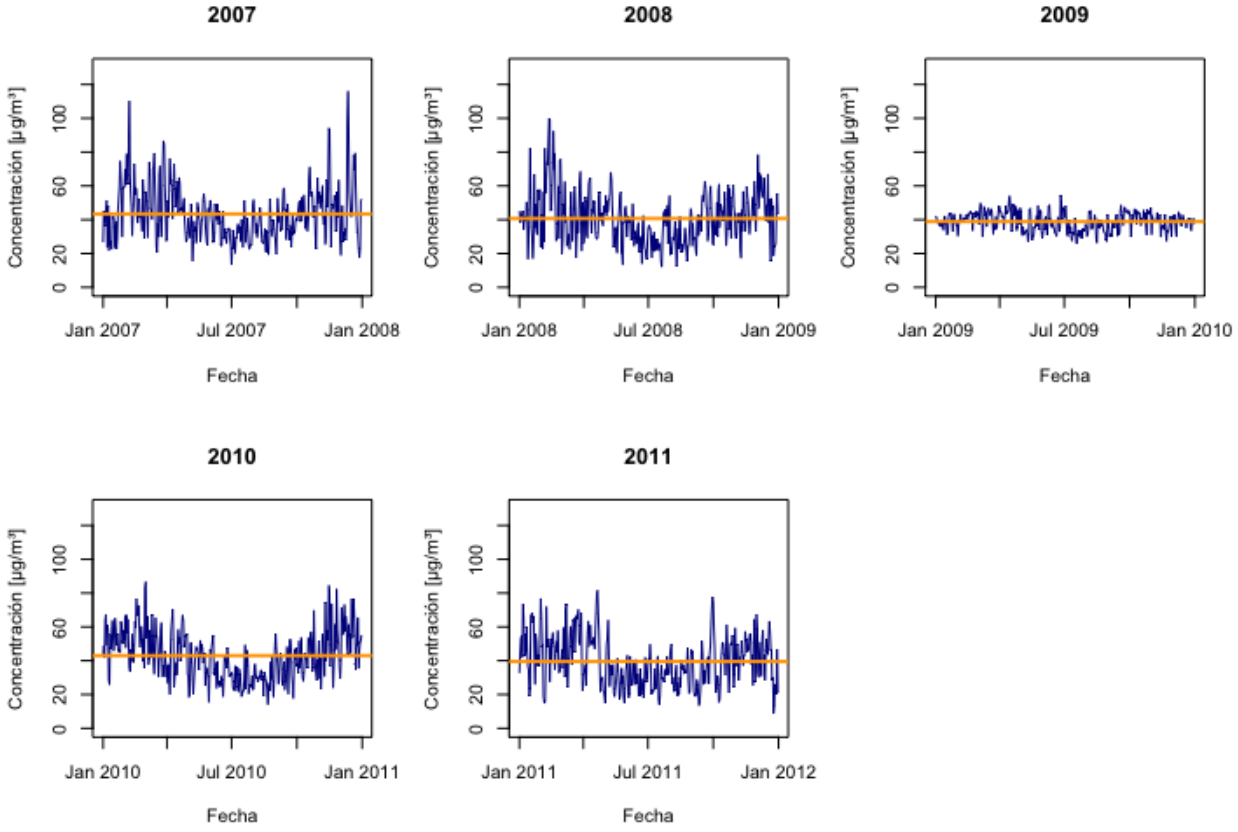


Figura 5: Datos medios diarios de la concentración de NO_2 en el periodo de estudio. Línea horizontal de color naranja representada a la altura de la media calculada en la Tabla 2 en cada año.

En las gráficas de la Figura 5 se puede ver claramente que los datos oscilan en torno a los valores medios de cada año, estando por encima en los meses de invierno y por debajo en los meses de verano.

Este aumento de la concentración del NO_2 en la atmósfera de Londres durante los meses de invierno y decrecimiento durante los meses de verano se puede explicar por varios factores relacionados con las condiciones climáticas y la actividad humana:

- Condiciones meteorológicas: Durante el verano, el aumento de la temperatura y la mayor actividad de los vientos pueden contribuir en la dispersión y dilución de los contaminantes atmosféricos y reducir su concentración en el aire.
- Actividad humana: En Londres la actividad humana y la densidad de población tienden a ser más altas durante el invierno por motivos laborales o educacionales. El mayor uso de vehículos en invierno debido a las condiciones climáticas supone un gran aumento en la concentración de NO_2. Durante esta temporada también se utilizan más sistemas de calefacción, lo que puede aumentar las emisiones de NO_2.
- Inversión térmica: Esto ocurre cuando una capa de aire cálido se forma sobre una capa de aire frío, lo que hace que los contaminantes se queden atrapados cerca del suelo. Esto lleva

a un aumento de la concentración de NO_2. En invierno las inversiones térmicas son más comunes, especialmente en áreas urbanas como Londres (García *et al.*, 2012).

Dado que las trayectorias de las concentraciones del NO_2 en la atmósfera observadas en la Figura 5 parecen seguir un patrón sinusoidal pero con ruido, es probable que múltiples de estos factores estén interactuando para dar como resultado estas fluctuaciones.

- Que el valor del rango intercuartílico sea menor en 2009 significa que los datos oscilan menos en torno al valor central. Esto se puede observar con claridad en la Figura 4.

- Los valores bajos del índice de Yule-Kendall indican también una distribución equilibrada de los valores en torno al valor central. Si el valor es negativo, los valores tienden a ser ligeramente más bajos que el valor central, mientras que si es positivo tienden a ser ligeramente más altos. Es decir, un valor positivo supone que la distancia a la mediana será mayor desde el cuartil superior que desde el cuartil inferior.

Se presenta ahora un diagrama de cajas y bigotes en la Figura 6 con una caja por año representando todos los valores de todas las regiones.

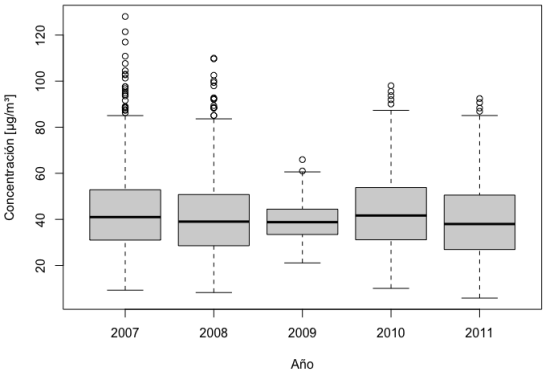


Figura 6: Evolución anual de las concentraciones de NO_2 en el periodo de estudio.

Con este diagrama se pueden sacar conclusiones muy similares a las anteriores:

- La tendencia es decreciente con un pico en 2010.
- Los datos están similarmente distribuidos por encima y por debajo de la mediana.
- La dispersión de los datos es menor en 2009.
- El valor de la mediana está en torno a los 40 $\frac{\mu g}{m^3}$.

Además se puede hacer una observación con respecto a los valores atípicos. El número de valores atípicos disminuye a medida que avanzamos en la serie temporal. Esto se puede deber, por ejemplo, a una mayor consistencia en las mediciones a lo largo del tiempo por una mejora en la precisión o calidad de los datos recopilados o a una estabilidad en el proceso. La estabilidad en el proceso se refiere a que el proceso está evolucionando de manera más predecible, sin fluctuaciones extremas

o inesperadas. Esta estabilidad se puede deber a un incremento en las regulaciones atmosféricas con respecto a emisión de contaminantes o a una mayor conciencia pública con respecto a la contaminación, entre otros.

Se representa ahora en la Figura 7 un diagrama de cajas y bigotes de la evolución mensual incluyendo los datos de las tres regiones.

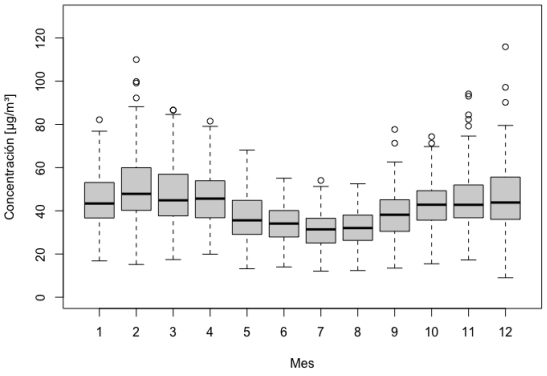
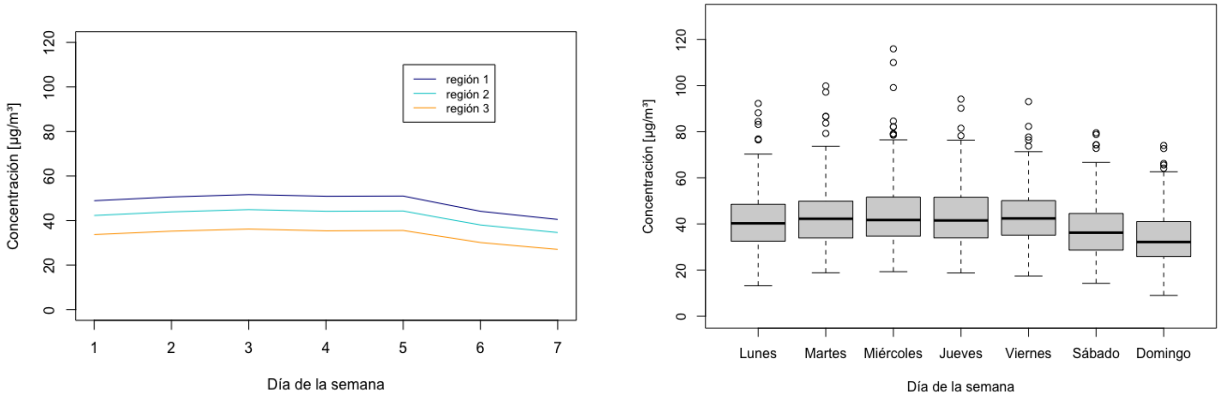


Figura 7: Evolución mensual de las concentraciones de NO_2 en el período de estudio.

En este diagrama también se ve reflejada la oscilación anual entorno al valor $40\frac{\mu g}{m^3}$, siendo más bajos los valores en los meses de verano y más altos en los meses de invierno.

Ahora se representan en la Figura 8 los datos por semana. Para ello se hace la media por regiones de los valores de todos los lunes, los martes, etc. y se representan los datos en función de los días de la semana. También se representa un diagrama de cajas y bigotes con todos los valores en función de los días de la semana.



(a) (b)

Figura 8: Evolución semanal en el periodo de estudio. (a) Valores de la concentración media de NO_2 en función del día de la semana a lo largo del periodo de estudio. (b) Diagrama de cajas y bigotes con todos los valores a lo largo del periodo de estudio en función del día de la semana.

En estos dos gráficos de la Figura 8 se ve cómo existe una clara tendencia a lo largo de los días de la semana. Durante los días laborables el valor de la concentración del NO_2 es constante mientras que el fin de semana decrece en gran medida.

Este decrecimiento se debe en gran medida al descenso o cese de la actividad industrial y el uso de vehículos, que son dos de los grandes emisores del contaminante NO_2.

En resumen, del análisis estadístico se pueden extraer las siguientes conclusiones:

- Se tiene una tendencia general de los datos linealmente decreciente, pero existe un pico de crecimiento en 2010.
- Existe un ciclo anual; los valores en los meses de verano son menores que el valor medio y los valores en los meses de invierno son mayores que el valor medio.
- Existe un ciclo semanal; los valores durante los días laborables mantienen un valor constante y cuando llega el fin de semana hay un decrecimiento de concentración de NO_2 en la atmósfera.

3.1.1. Predicción con el algoritmo Prophet. Primer acercamiento

A raíz del análisis que se acaba de exponer, se ha decidido realizar los ajustes y predicciones para una sola región de las proporcionadas para el estudio. Se han utilizado los datos de la región uno.

El algoritmo Prophet tiene muchos parámetros establecidos por defecto, pero ofrece la posibilidad de modificarlos para mejorar la efectividad del modelo de predicción.

Como ya hemos detectado, nuestra serie temporal muestra una periodicidad anual y semanal, luego se han establecido los parámetros yearly_seasonality = TRUE, weekly_seasonality = TRUE. También permite establecer los días festivos, que normalmente tienen un impacto considerable en los patrones humanos que influyen en la emisión de NO_2 a la atmósfera. Se ha creado un dataframe con las vacaciones que queramos incluir y sus correspondientes fechas pasadas y futuras dentro del rango utilizado para las predicciones. El dataframe es como el que se muestra en la Figura 9.

	holiday	ds	lower_window	upper_window
1	newYear	2007-01-01	1	1
2	newYear	2008-01-01	1	1
3	newYear	2009-01-01	1	1
4	newYear	2010-01-01	1	1
5	newYear	2011-01-03	1	1
6	easter	2007-04-06	1	1
7	easter	2007-04-07	1	1
8	easter	2007-04-08	1	1
9	easter	2007-04-09	1	1
10	easter	2008-03-21	1	1

Figura 9: Parte inicial del dataframe con las vacaciones de Londres en el periodo de estudio y sus respectivas fechas.

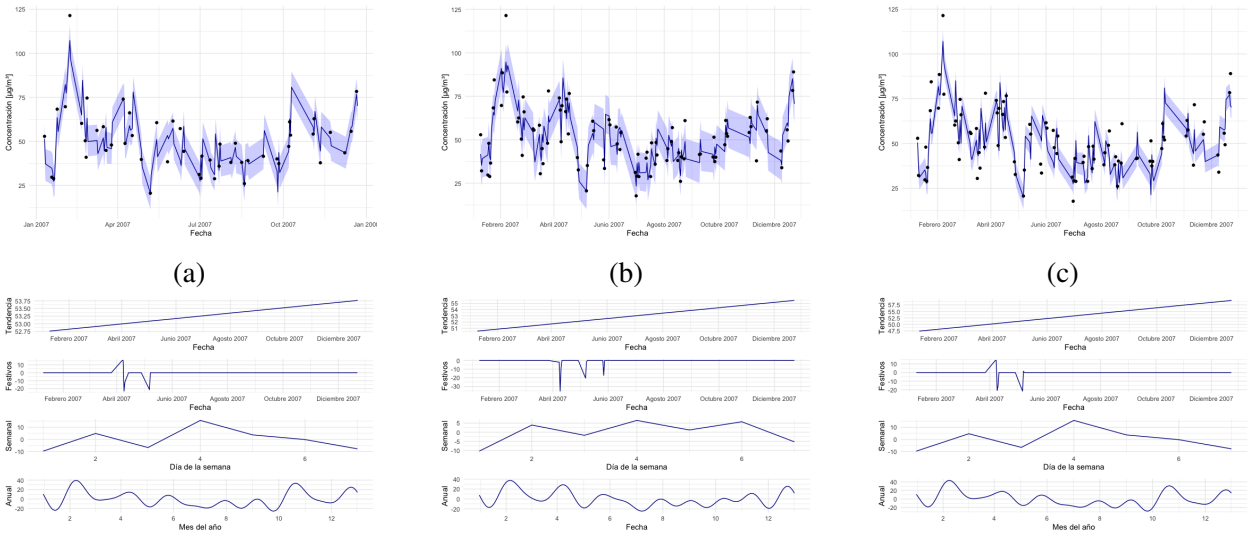
En este dataframe se incluyen los días festivos que se repiten todos los años en Londres, que son 'New year's day', 'Easter' (incluyendo 'Good Friday' y 'Easter Monday'), 'Early May Bank Holidays', 'Spring Bank Holidays', 'Late summer Bank Holidays', 'Christmas Day' y 'Boxing Day'. También se incluye la boda real celebrada el 29 de abril de 2011, que fue fiesta nacional.

Las columnas de *lower_window* y *upper_window* indican el número de días hacia el pasado y hacia el futuro partiendo desde cada día festivo que el algoritmo debiera considerar para detectar patrones y tendencias relacionadas con el día festivo. En este caso se añaden los días anterior y posterior.

Del análisis visual de las gráficas de todos los acercamientos presentadas en el Anexo A podemos sacar las siguientes conclusiones:

- Si se ajusta el modelo con pocos valores muy separados en el tiempo, las componentes del modelo no son visualmente correctas. No son suficientes datos para que el modelo identifique una periodicidad anual ni semanal específica. Esto se puede observar en la Figura 39.
- Si se ajusta el modelo con pocos valores, la componente de los días festivos y eventos marca picos de influencia en el cambio de año. Hay un cambio de tendencia cada año y el algoritmo identifica las vacaciones de Navidad como factor influyente en este cambio de tendencia. Esto también se puede observar en la Figura 39.
- En ocasiones las componentes son completamente erróneas o se alejan visualmente bastante de la realidad. Esto se ha asociado a una mala selección de los días aleatorios con los que se ajusta el modelo. Esto por ejemplo pasa en la Figura 40.
- En general, cuantos más datos utilicemos para ajustar el modelo, más se acercan los resultados de las componentes a la realidad.
- La componente de la tendencia general de la serie no está bien definida en ningún caso, no es suficiente el número de valores empleado para ajustar de manera correcta el modelo.
- Los valores específicos predichos siguen bien la trayectoria de la serie original porque se predicen los valores con los que se ajusta el modelo y los de días consecutivos, por lo cual la predicción es similar al valor real. Esto se ve en la fila de arriba de cada una de las gráficas.
- Las componentes de los modelos ajustados en los pasos uno y tres son muy similares, si no idénticas entre sí. Esto ocurre porque en el paso tres se ajusta el modelo con los valores predichos en el paso uno. La componente que más cambia entre los pasos uno y tres es la tendencia general de la serie, porque en el paso tres se utilizan más datos para ajustar el modelo que en el paso uno.

Del primer acercamiento se mostrarán las predicciones de los valores del NO_2 seleccionados únicamente de entre los valores de 2007 y algunas predicciones que sean de especial interés o representativas. Del segundo acercamiento se mostrarán todas las predicciones. El resto de las predicciones realizadas se muestran en el Anexo A.



(a) (b) (c)

(d) (e) (f)

Figura 10: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de 2007. (a) Paso uno. (b) Paso dos. (c) Paso tres. (d) Componentes paso uno. (e) Componentes paso dos. (f) Componentes paso tres.

En la fila de arriba de la Figura 10 se representan los valores predichos en azul oscuro con formato línea junto con el rango de error en azul claro y con puntos negros los valores con los que se ajusta el modelo. En la fila de abajo se representan las diferentes componentes del modelo.

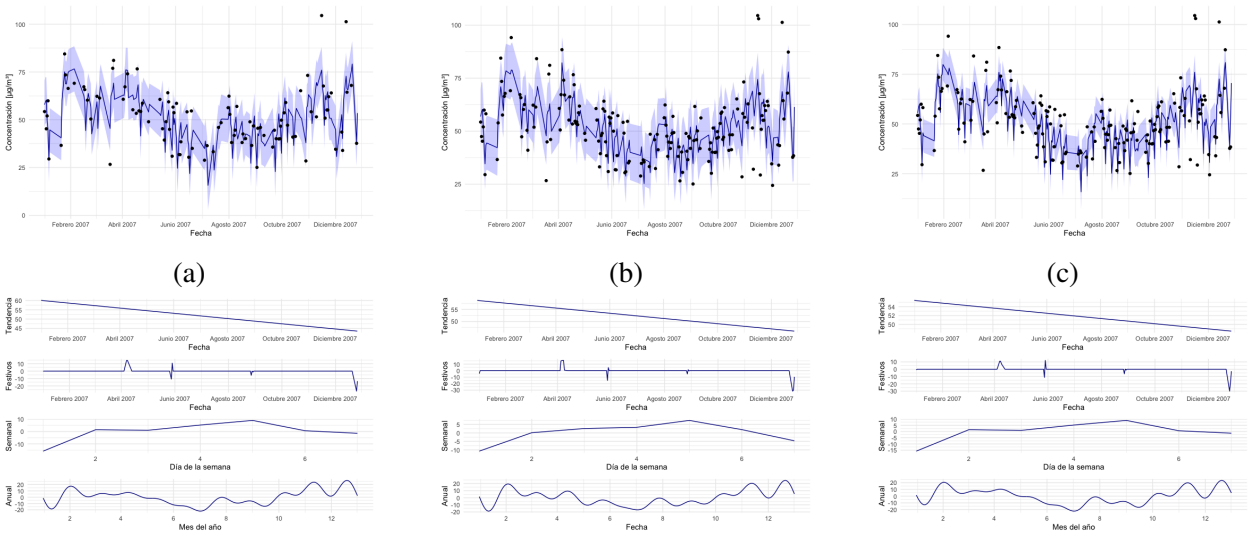
Las componentes del modelo son calculadas por el algoritmo en función de los valores con los que ajustes el modelo. Las componentes representadas son de arriba a abajo:

- Tendencia general de la serie
- Impacto de los días festivos en la predicción
- Tendencia semanal
- Tendencia anual

Hay que tener en cuenta que en las gráficas de la tendencia semanal la numeración de los días empieza desde el domingo, es decir, el domingo es el día uno, el lunes es el día dos, etc.

La primera columna de la Figura 36 corresponde al paso uno explicado en la sección 2 Materiales y Métodos, la segunda al paso dos y la tercera al paso tres.

Como se puede ver en el paso uno de la Figura 36, la tendencia semanal está bastante mal definida y la tendencia anual ya muestra signos de que en los meses de invierno la concentración es mayor que en los meses de verano. La tendencia general de la serie se muestra creciente cuando debería ser decreciente. La selección de datos aleatorios no es lo suficientemente representativa como para definir adecuadamente las componentes. En el paso dos la tendencia semanal se acerca más a la realidad. En el paso tres los resultados son muy similares a las del paso uno porque se ajusta el modelo con las componentes calculadas en el paso uno.

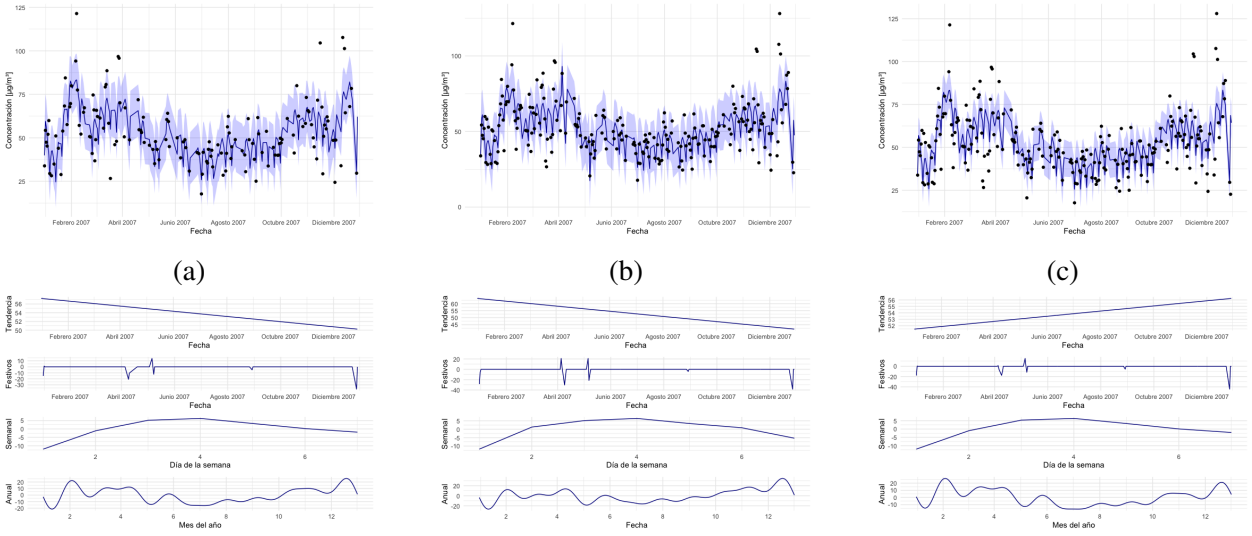


(a) (b) (c)

(d) (e) (f)

Figura 11: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de 2007. (a) Paso uno. (b) Paso dos. (c) Paso tres. (d) Componentes paso uno. (e) Componentes paso dos. (f) Componentes paso tres.

Se puede ver en la Figura 11 cómo al incluir más datos la tendencia general de la serie ya se define correctamente descendiente. La tendencia semanal y anual de la serie también se acercan un poco más a como hemos descrito que es la realidad.

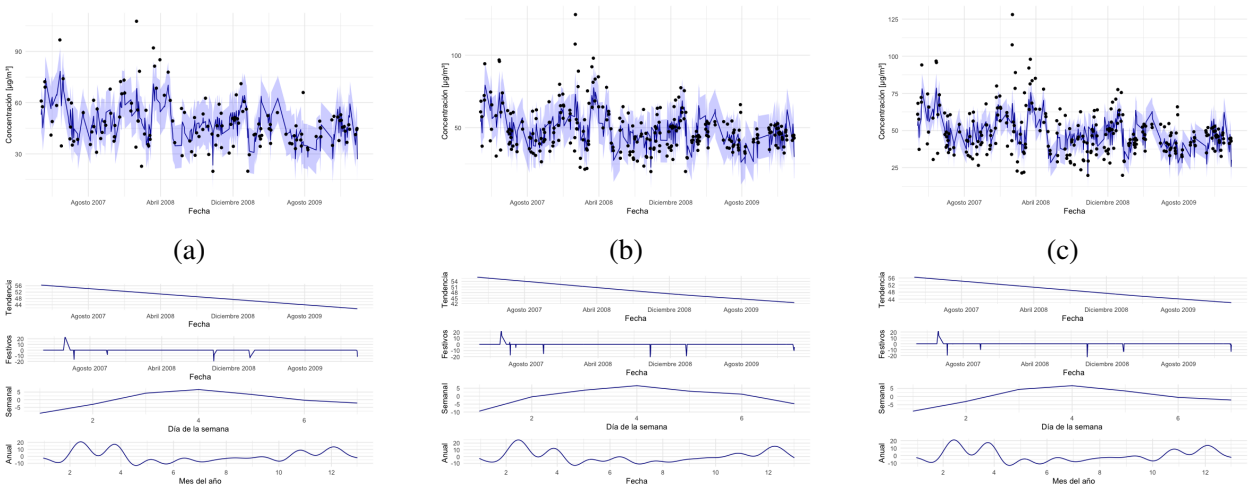


(a) (b) (c)

(d) (e) (f)

Figura 12: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de 2007. (a) Paso uno. (b) Paso dos. (c) Paso tres. (d) Componentes paso uno. (e) Componentes paso dos. (f) Componentes paso tres.

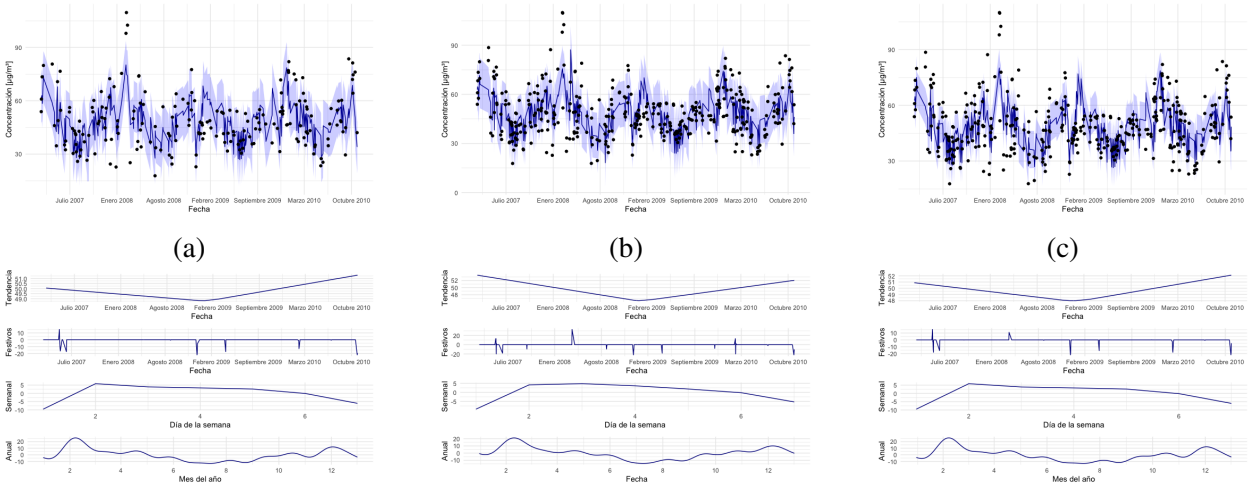
En la Figura 12 se observa cómo las curvas se van asemejando más a los datos reales. La trayectoria general de la curva es similar a la que hemos observado durante el análisis descriptivo.



(a) (b) (c)

(d) (e) (f)

Figura 13: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de 2007, 2008 y 2009. (a) Paso uno. (b) Paso dos. (c) Paso tres. (d) Componentes paso uno. (e) Componentes paso dos. (f) Componentes paso tres.



(a) (b) (c)

(d) (e) (f)

Figura 14: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de 2007, 2008, 2009 y 2010. De arriba a abajo predicciones y componentes. (a) paso uno. (b) paso dos. (c) paso tres. (d) Componentes paso uno. (e) Componentes paso dos. (f) Componentes paso tres.

En la figura 14, en los pasos dos y tres, se observa cómo la componente de la tendencia general de la serie se aproxima bastante al comportamiento general de la serie que se ha observado durante el análisis estadístico de la misma.

Se presentan en la Tabla **??** los errores absolutos medios de las predicciones con respecto a los valores reales de la serie de datos utilizada en el estudio. Se presentan en función del número de

días aleatorizados que se utilizan para ajustar el modelo, especificando el paso al que pertenece la predicción sobre la que se calcula el error cuadrático medio.

Tabla 3: Errores absolutos medios de las predicciones realizadas utilizando los modelos ajustados. Paso representa el paso en el que se realiza la predicción. El año representa hasta qué año se incluye, partiendo desde 2007, para realizar la selección de días aleatorizados.

		Número de días aleatorizados			
	Paso	50	100	150	200
2007	1	8,737	8,778	9,691	
	2	8,200	8,708	9,516	
	3	10,123	9,694	9,884	
2008	1	6,802	10,539	11,579	10,631
	2	8,185	10,090	10,226	10,129
	3	8,819	10,827	11,624	10,881
2009	1	10,764	8,223	9,207	8,314
	2	8,062	8,415	9,316	8,435
	3	11,512	8,945	9,881	8,856
2010	1	9,097	9,215	8,594	9,154
	2	9,403	8,745	8,362	9,016
	3	10,446	9,477	8,981	9,378
2011	1	6,695	8,848	9,365	9,200
	2	6,816	8,247	8,988	8,959
	3	7,721	9,081	9,674	9,400

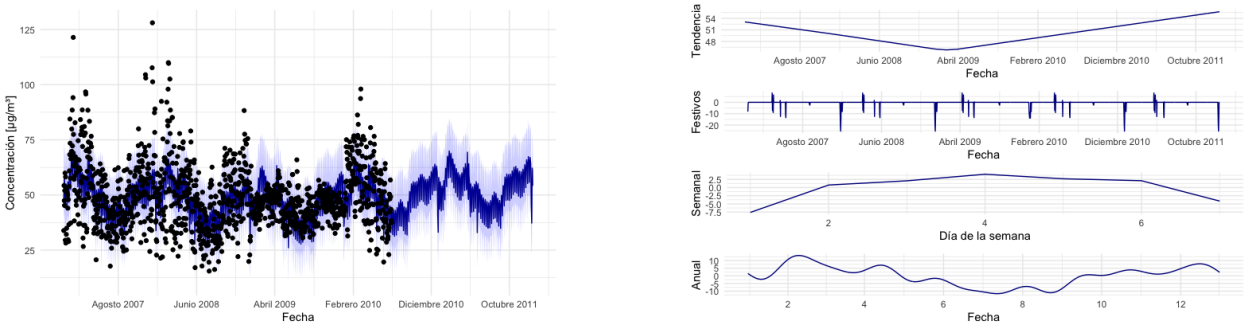
Cuanto menos error absoluto medio, mejor hecha está la predicción. Se puede ver cómo en general la mejor predicción es en el paso dos. Esto significa que el algoritmo mejora cuando el modelo se ajusta con una sucesión de datos en la que hay datos consecutivos. El paso tres tiene en general la peor predicción, ya que 'arrastra' el error de la predicción del paso uno.

3.1.2. Predicción con el algoritmo Prophet. Segundo acercamiento

Ahora se utilizan datos consecutivos para ajustar el modelo.

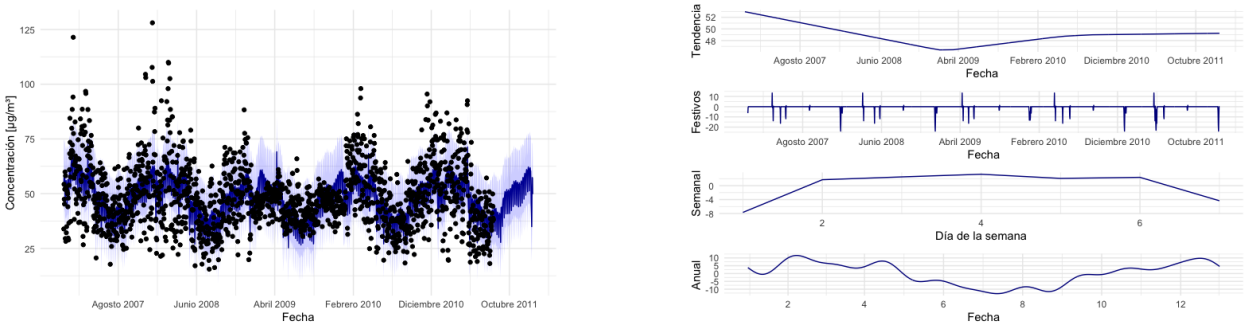
En este caso la predicción o pronóstico ajusta de manera muy satisfactoria tanto la componente de la estacionalidad semanal como el de la estacionalidad anual. En la componente de la tendencia se debería observar una tendencia de la misma forma a la representada en la Figura 15 pero con un cambio entre 2010 y 2011 a tendencia decreciente.

En este caso no se ve este cambio. Al usar los datos hasta julio de 2010 no se observa la tendencia decreciente que hay después del pico de valores de 2010. El algoritmo no es capaz de detectarlo.



(a) (b)

Figura 15: Modelo ajustado con los datos de la región 1 del NO_2 desde el 2007-01-01 hasta el 2010-07-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.



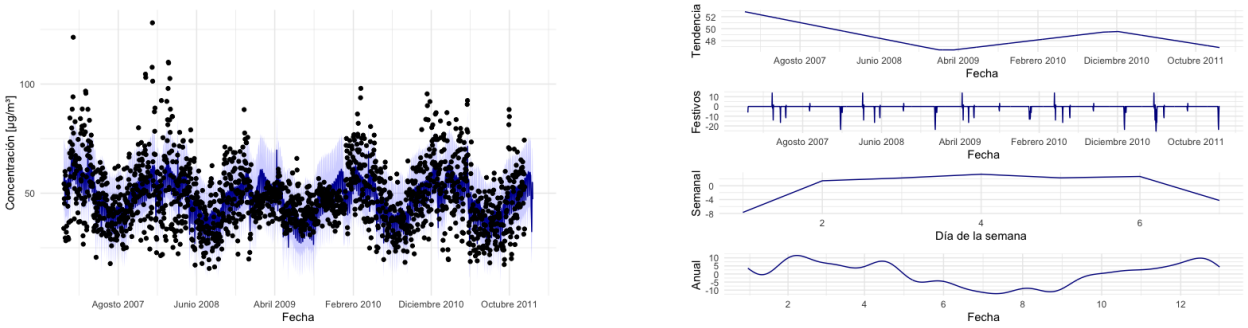
(a) (b)

Figura 16: Modelo ajustado con los datos de la región 1 del NO_2 desde el 2007-01-01 hasta el 2011-08-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

Incluyendo un año más de valores, las estacionalidades semanal y anual se definen casi perfectas, mientras que la tendencia general de la serie se aplana a partir de 2011.

Todavía no se consigue que la tendencia general de la serie se modele de manera completamente correcta. Si se hiciera una predicción de valores futuros desconocidos, estos seguirían esta tendencia plana que tiene la serie predicha en 2011. Por lo tanto, no se ajustarían con el decrecimiento general que debiera experimentar la serie a lo largo de los años.

Para conseguir que la tendencia general de la serie quede ajustada a la realidad completa de los datos, se ajusta el modelo con prácticamente todos los datos que se poseen.



(a) (b)

Figura 17: Modelo ajustado con los datos de la región 1 del NO_2 desde el 2007-01-01 hasta el 2011-12-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

Con este conjunto de datos para ajustar el modelo se puede ver cómo la componente de la tendencia queda correctamente ajustada a la realidad y también se consigue la mejor aproximación de las componentes de las estacionalidades semanal y anual.

Se muestra en la Tabla 4 los errores absolutos medios, que se utilizan para medir la bondad de los modelos de ajuste y predicción, en función del paso del proceso que se sigue en el segundo acercamiento.

Tabla 4: Error absoluto medio de los valores predichos por el algoritmo Prophet durante el acercamiento dos con respecto a los valores reales.

paso	EAM
1	9.975
2	9.523
3	9.484

Es claro que la predicción es mejor cuantos más datos se utilicen para ajustar el modelo.

3.2. Análisis estadístico y predicción del ozono u O_3

Se ha seguido el mismo proceso que con las concentraciones del NO_2 y el análisis es bastante similar, por lo que se presentan únicamente las gráficas más representativas.

Se representan en la Figura 18 los datos de la concentración a lo largo del tiempo con un color diferente para cada región.

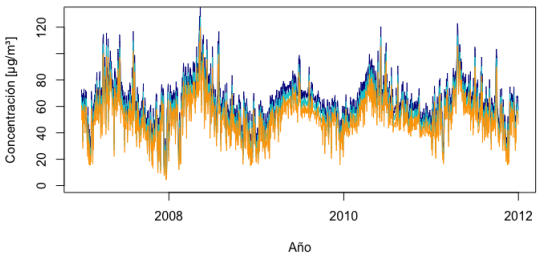


Figura 18: Concentración de O_3 en el periodo de estudio en tres regiones: Región 1 (azul oscuro), Región 2 (azul claro), Región 3 (naranja).

En la Tabla 5 se muestran los principales estadísticos de estudio.

Tabla 5: Tabla con estadísticos calculados para analizar la tendencia general de la serie.

año	media	mediana	rango intercuartílico	índice Yule - Kendall
2007	54.775	55.231	23.984	-0.038
2008	58.740	58.101	26.192	-0.014
2009	55.862	57.158	17.543	-0.081
2010	59.003	56.861	17.115	0.119
2011	57.782	56.710	19.894	0.086

Los valores de esta tabla permiten llegar a las siguientes conclusiones:

- Que los valores de la media y de la mediana sean similares entre sí indica que hay una distribución equilibrada de valores tanto por encima como por debajo del valor central. Además, se puede decir que los valores están similarmente distribuidos por encima y por debajo de la mediana y que, o no hay gran cantidad de valores atípicos, o están asimismo distribuidos equilibradamente por encima y por debajo de la mediana, ya que en otro caso la media se vería sesgada hacia un extremo. Se puede utilizar tanto la media como la mediana para describir la dirección general que sigue nuestra serie temporal.

- Con los valores de la media se puede ver cómo el valor de la concentración de O_3 sigue una tendencia general creciente escalonada bianualmente. Es decir, en 2009 el valor de la media es mayor que en 2007, en 2010 es mayor que en 2008 y en 2011 es mayor que en 2009, pero después de cada año par hay una caída en la concentración. Se necesitaría una serie más larga para confirmar la tendencia.

Se representa en la Figura 19 un diagrama de cajas y bigotes de las medias mensuales de todos los años incluyendo los datos de las tres regiones.

Este patrón anual de la concentración de O_3 en la atmósfera de Londres se puede explicar principalmente por dos factores: la radiación solar y las condiciones meteorológicas.

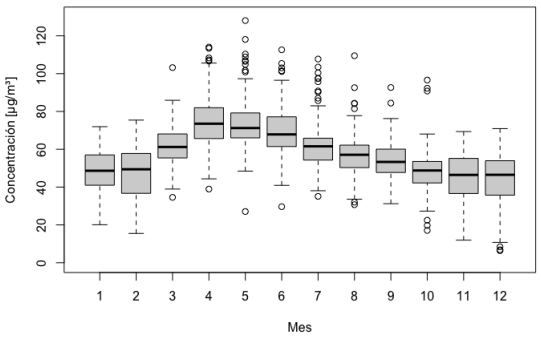


Figura 19: Diagrama de cajas y bigotes para representar la evolución de la serie a lo largo de los meses en un año. Se utilizan los valores de los cinco años por cada mes.

- Radiación solar: La radiación solar aumenta durante el verano porque el sol tiene una posición más alta en el cielo y además los días son más largos. La mayor radiación ultravioleta que proviene del sol durante el verano interacciona con los contaminantes atmosféricos (óxidos de nitrógeno NO_x, compuestos orgánicos volátiles, etc.) y provoca reacciones químicas que conducen a la formación de O_3.
- Condiciones meteorológicas: Las altas temperaturas y la mayor actividad de los vientos pueden contribuir en la dispersión y dilución de los contaminantes atmosféricos y reducir su concentración en el aire, pero la mayor estabilidad de las condiciones meteorológicas y la falta de lluvia en verano contribuyen a la acumulación de O_3, ya que impiden una dispersión eficiente tanto de los contaminantes como del O_3 generado.

En este diagrama se observa que los valores más altos se encuentran en los meses de abril y mayo, empiezan a decrecer a lo largo de los meses de verano y son más bajos en los meses de invierno.

Esto probablemente se deba a la reducción de la contaminación de óxidos de nitrógeno NO_x y compuestos orgánicos volátiles durante los meses de verano. A pesar de haber mayor radiación solar en los meses de verano, también hay menos contaminantes con los que interactuar para producir ozono. En los meses de abril y mayo la radiación solar aumenta y la producción de contaminantes se mantiene prácticamente constante porque son meses laborales y escolares, lo que supone un uso constante de transportes e industria.

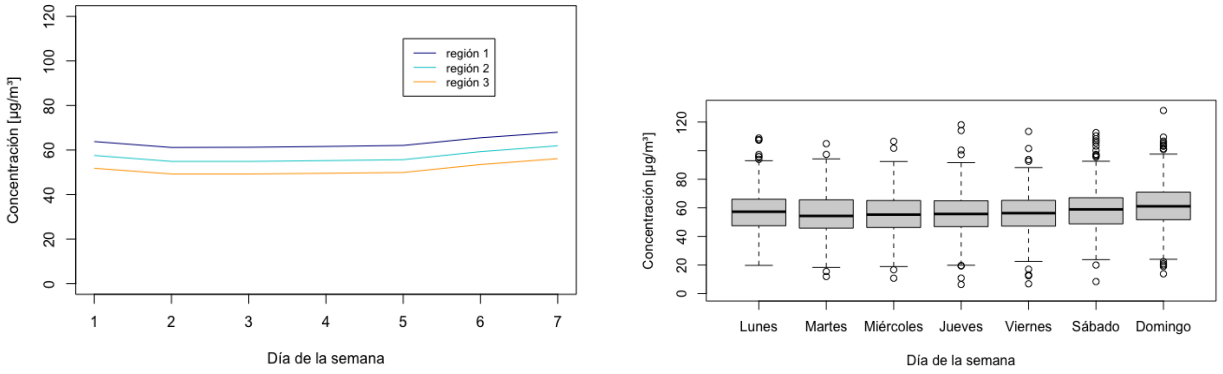
También se puede ver una gran dispersión de los valores en los meses de verano, siendo valores más constantes durante los meses de invierno. Esto significa que a lo largo de los años la concentración de O_3 durante los meses de invierno no varía mucho, mientras que en los meses de verano las concentraciones son más diferentes de año a año, siendo en su mayoría mayores que la mediana.

Esto se debe a que hay días en verano en los que la contaminación es lo suficientemente alta como para que la producción de O_3 por la gran cantidad de radiación solar sea igual, o incluso mayor, que en los meses de abril y mayo.

- Los valores del rango intercuartílico son bastante constantes, indicando que la oscilación de los valores alrededor de la mediana es de una magnitud similar a lo largo de los años.

- Los valores bajos del índice de Yule-Kendall pueden explicarse de igual manera que los valores explicados en la Tabla 2.

Se representan en la Figura 20 los datos de media por semana y un diagrama de cajas y bigotes con todos los valores en función de los días de la semana.



(a) (b)

Figura 20: Tendencia semanal de la serie temporal. (a) Valores de la concentración media de O_3 en función del día de la semana a lo largo del periodo de estudio. (b) Diagrama de cajas y bigotes con todos los valores a lo largo del periodo de estudio en función del día de la semana.

En estos dos gráficos se ve cómo existe una tendencia a lo largo de los días de la semana. Durante los días laborables el valor de la concentración del O_3 es constante mientras que el fin de semana aumenta mínimamente.

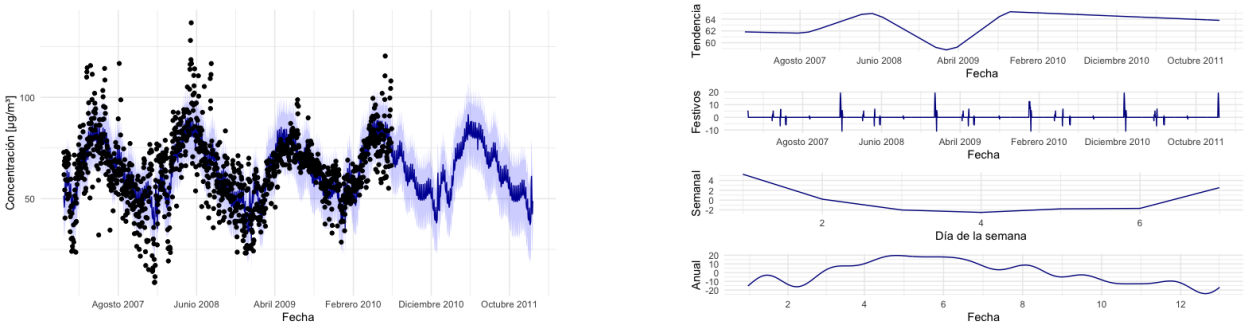
En lugares con altos niveles de NO, una vez formado el O_3 se consume rápidamente por la oxidación del NO al NO_2. Esto explica esta periodicidad semanal. Los fines de semana se reducen el tráfico y la industria, que son los principales productores de NO, luego hay menos concentración de NO en la atmósfera y por lo tanto el O_3 no se ve tan consumido y su concentración aumenta.

En resumen, del análisis estadístico se pueden extraer las siguientes conclusiones:

- Existe una tendencia general de los datos linealmente creciente escalonada bianualmente, aunque se necesitaría un periodo de medidas más amplio para ver si esta tendencia se confirma.
- Hay una periodicidad anual; los valores en los meses de verano son mayores que el valor medio y los valores en los meses de invierno son menores que el valor medio.
- Se observa un ciclo semanal; los valores durante los días laborables mantienen un valor constante y cuando llega el fin de semana hay un aumento de concentración de O_3 en la atmósfera.

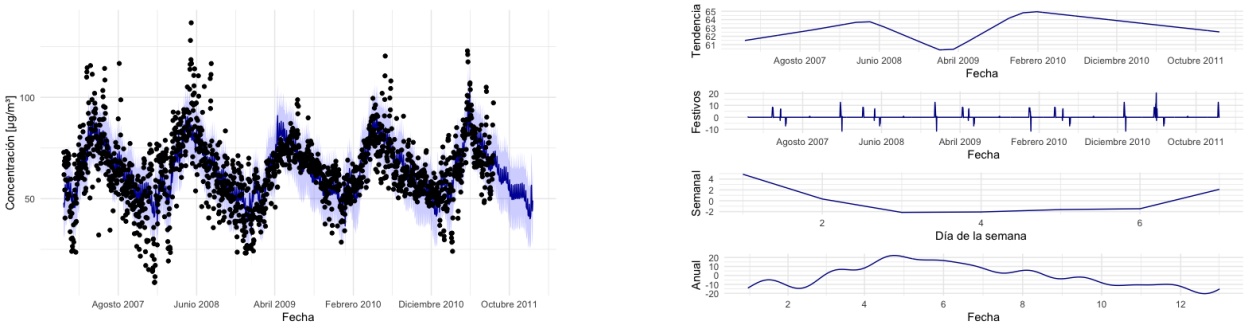
3.2.1. Predicción con el algoritmo Prophet. Segundo acercamiento

En este caso sólo se presentan los resultados del segundo acercamiento porque son los que mejor calculan las componentes del modelo. El resto de predicciones y estadísticos se presentan en el Anexo B. Por lo tanto, se utilizan datos consecutivos para ajustar el modelo.



(a) (b)

Figura 21: Modelo ajustado con los datos de la región 1 del O_3 desde el 2007-01-01 hasta el 2010-07-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

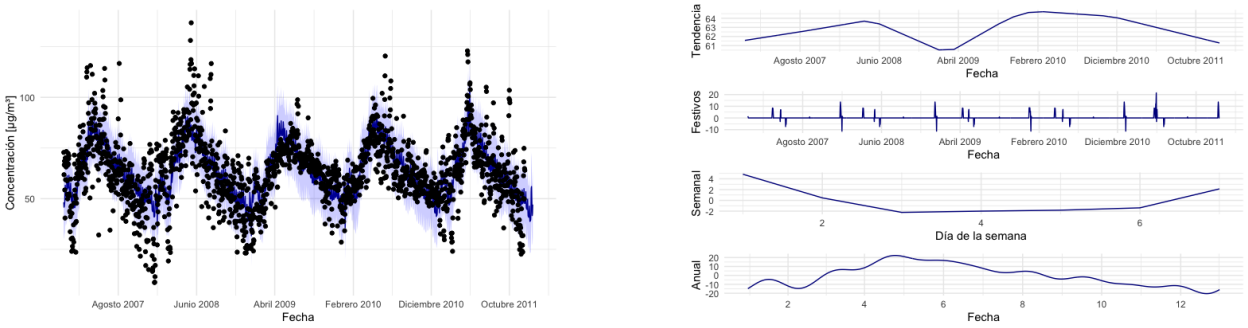


(a) (b)

Figura 22: Modelo ajustado con los datos de la región 1 del O_3 desde el 2007-01-01 hasta el 2011-08-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

Incluyendo un año más de valores, la tendencia general de la serie en la Figura 22 se define visualmente perfecta. En particular, se puede ver en la Figura 22b que las estacionalidades semanal y anual casi se definen por completo, teniendo el sábado menos concentración que el domingo, lo cual se ha visto en el análisis inicial.

Con este conjunto de datos para ajustar el modelo se puede ver en la Figura 23 cómo la componente de la tendencia queda correctamente ajustada a la realidad y también se consigue la mejor aproximación de las componentes de las estacionalidades.



(a) (b)

Figura 23: Modelo ajustado con los datos de la región 1 del O_3 desde el 2007-01-01 hasta el 2011-12-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

Se muestra en la Tabla 6 los errores absolutos medios, que se utilizan para medir la bondad de los modelos de ajuste y predicción, en función del paso del proceso que se sigue en el segundo acercamiento.

Tabla 6: Error absoluto medio de los valores predichos por el algoritmo Prophet durante el acercamiento dos con respecto a los valores reales.

paso	EAM
1	9.131
2	9.027
3	9.020

De nuevo, es claro que la predicción es mejor cuantos más datos se utilicen para ajustar el modelo, igual que con las predicciones del NO_2.

3.3. Análisis estadístico y predicción de las PM_{10}

Se ha seguido el mismo procedimiento que anteriormente.

Se representan en la Figura 24 los datos de la concentración a lo largo del tiempo con un color diferente para cada región.

En la Tabla 7 se muestran los principales estadísticos de estudio.

Como primera observación, de la observación de la Figura 24, es claro que en este conjunto de datos hay menor diferencia entre los valores de las tres regiones, siguiendo las tres la misma trayectoria a lo largo del periodo de estudio.

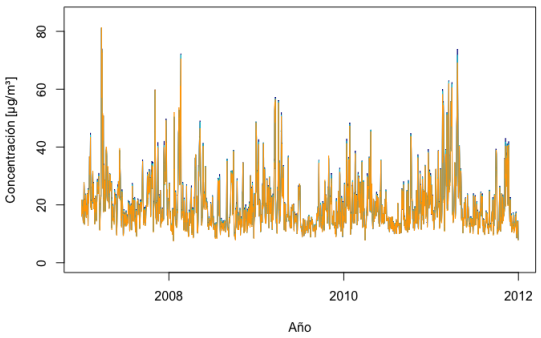


Figura 24: Concentración de PM_{10} en el periodo de estudio en tres regiones: Región 1 (azul oscuro), Región 2 (azul claro), Región 3 (naranja).

Tabla 7: Tabla con estadísticos calculados para analizar la tendencia general de la serie.

año	media	mediana	rango intercuartílico	índice Yule - Kendall
2007	22.674	20.546	9.117	0.178
2008	19.079	16.087	8.653	0.293
2009	18.616	16.290	8.178	0.260
2010	19.236	17.259	8.285	0.226
2011	20.968	17.321	10.290	0.285

Los valores de las medianas son menores que los valores de las medias y distan de estas una cantidad constante a lo largo de los años. Esto nos indica que hay mayor irregularidad de los datos por encima de la mediana. Esto se puede observar claramente en la Figura 24. Que el índice de Yule-Kendall sea positivo y bastante alejado de cero, también nos indica que hay una clara tendencia de los datos a estar por encima de la mediana.

Los valores de la media oscilan en torno a un valor de 20 $\frac{\mu g}{m^3}$, teniendo la serie una tendencia decreciente hasta 2009 y creciente desde entonces hasta 2011.

Se representa en la Figura 25 un diagrama de cajas y bigotes de las medias mensuales de todos los años incluyendo los datos de las tres regiones.

El patrón anual de la concentración de PM_{10} en la atmósfera de Londres observado en la Figura 25 recuerda al presentado en la Figura 7, siendo de igual manera sinusoidal, pero en el caso de las PM_{10} con una amplitud menor. Esto significa que la concentración de las PM_{10} oscila alrededor de un valor central en menor medida que la concentración del NO_2. Esto también se puede apreciar al ver en la Tabla 7 que los valores del rango intercuartílico son menores que los presentados en la Tabla 2.

Este patrón anual de la serie, al igual que el patrón de las concentraciones de NO_2, se puede explicar principalmente por dos factores: la actividad humana y la inversión térmica.

- Actividad humana: En Londres la actividad humana y la densidad de población tienden a ser más altas durante el invierno por motivos laborales o educacionales. El mayor uso de

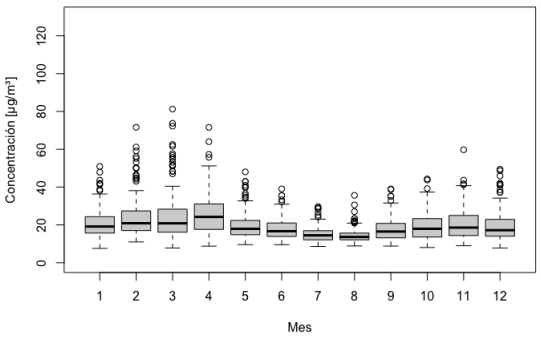


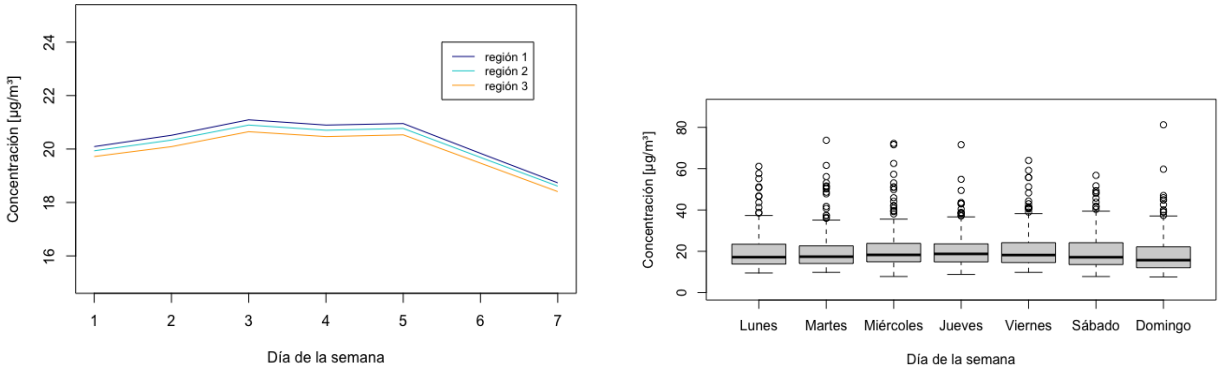
Figura 25: Diagrama de cajas y bigotes para representar la evolución de la serie a lo largo de los meses en un año. Se utilizan los valores de los cinco años por cada mes.

vehículos en invierno debido a las condiciones climáticas supone un gran aumento en la concentración de las PM_{10}. Durante esta temporada también se utilizan más sistemas de calefacción, lo que puede aumentar las emisiones de las PM_{10}.

- Inversión térmica: Esto ocurre cuando una capa de aire cálido se forma sobre una capa de aire frío, lo que hace que los contaminantes se queden atrapados cerca del suelo. Esto lleva a un aumento de la concentración de las PM_{10}. En invierno las inversiones térmicas son más comunes, especialmente en áreas urbanas como Londres.

En la Figura 25 también se puede ver la gran dispersión de los valores por encima de la mediana.

Se representan en la Figura 26 los datos de media por semana y un diagrama de cajas y bigotes con todos los valores en función de los días de la semana.



(a) (b)

Figura 26: Tendencia semanal de la serie temporal. (a) Valores de la concentración media de PM_{10} en función del día de la semana a lo largo del periodo de estudio. (b) Diagrama de cajas y bigotes con todos los valores a lo largo del periodo de estudio en función del día de la semana.

En estos dos gráficos se ve cómo existe una tendencia a lo largo de los días de la semana. Durante los días laborables el valor de la concentración del PM_{10} es constante mientras que el fin de semana decrece notablemente.

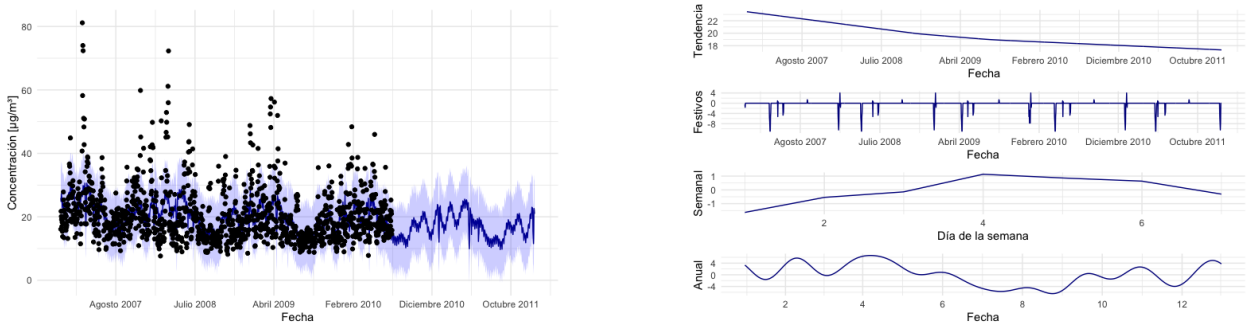
Este decrecimiento se debe en gran medida al descenso o cese de la actividad industrial y el uso de vehículos, que son dos de los grandes emisores del contaminante PM_{10}.

En resumen, del análisis estadístico se pueden extraer las siguientes conclusiones:

- El análisis de las PM_{10} es muy similar al del NO_2, teniendo el de las PM_{10} una mayor tendencia de los valores a estar por encima de la mediana y una oscilación de menor amplitud de estos en torno al valor de la mediana.
- Existe un ciclo anual; los valores en los meses de verano son menores que el valor medio y los valores en los meses de invierno son mayores que el valor medio.
- Existe un ciclo semanal; los valores durante los días laborables mantienen un valor constante y cuando llega el fin de semana hay un decrecimiento de concentración de PM_{10} en la atmósfera.

3.3.1. Predicción con el algoritmo Prophet. Segundo acercamiento

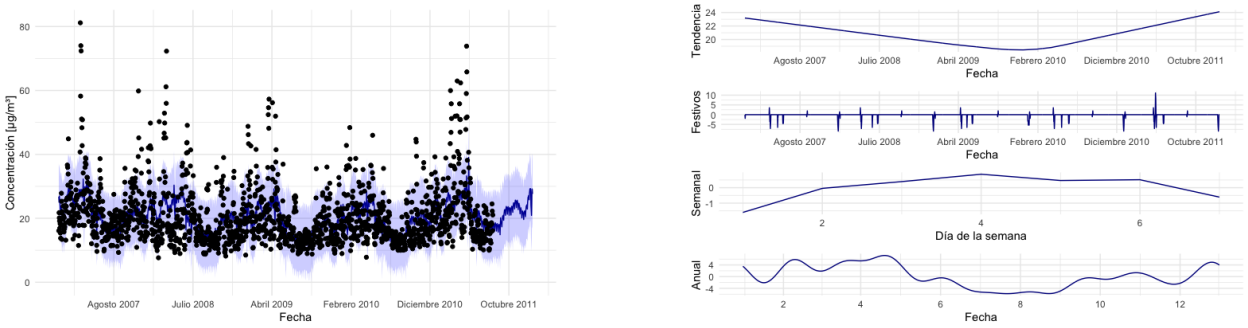
De nuevo, sólo se presentan los resultados del segundo acercamiento porque son los que mejor calculan las componentes del modelo. El resto de predicciones y estadísticos se presentan en el Anexo C.



(a) (b)

Figura 27: Modelo ajustado con los datos de la región 1 del PM_{10} desde el 2007-01-01 hasta el 2010-07-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

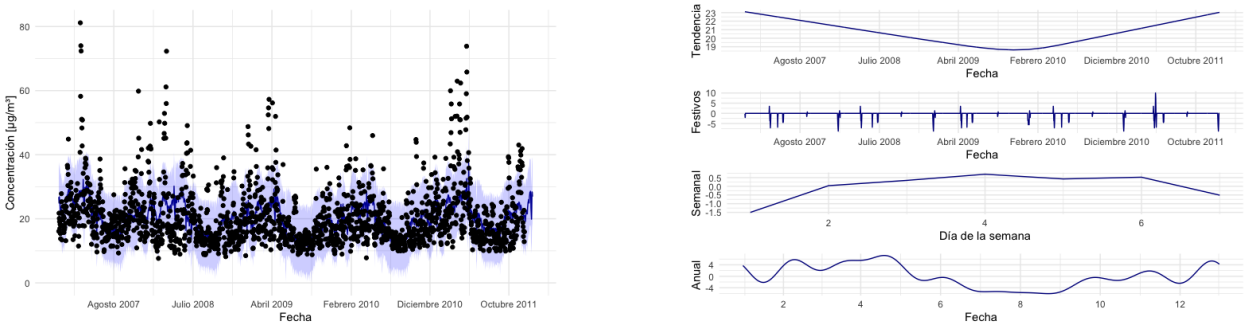
Es claro cómo la mejor predicción es la presentada en la Figura 29. Las tendencia general y semanal de la serie se definen en perfección. La componente anual refleja correctamente el patrón anual que siguen los datos.



(a) (b)

Figura 28: Modelo ajustado con los datos de la región 1 del PM_{10} desde el 2007-01-01 hasta el 2011-08-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

Si se hiciera una predicción de valores más allá de los valores conocidos, probablemente la tendencia seguiría siendo creciente, como es al final del modelo ajustado. No se puede saber si esto se ajusta a la realidad o no sin tener más información sobre los valores de la serie a partir de final del año 2011.



(a) (b)

Figura 29: Modelo ajustado con los datos de la región 1 del PM_{10} desde el 2007-01-01 hasta el 2011-12-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

Se muestra en la Tabla 8 los errores absolutos medios, que se utilizan para medir la bondad de los modelos de ajuste y predicción, en función del paso del proceso que se sigue en el segundo acercamiento.

En este caso los errores absolutos medios son menores que los presentados en la Tabla 4 y en la Tabla 6. La menor dispersión de los datos, o el menor número de datos atípicos, ayudan a que la predicción sea más exacta.

Tabla 8: Error absoluto medio de los valores predichos por el algoritmo Prophet durante el acercamiento dos con respecto a los valores reales.

paso	EAM
1	5.816
2	5.930
3	5.870

Los tres valores presentados en la Tabla 8 son muy parecidos, dando a entender que las tres predicciones son de una calidad similar con respecto a los valores con los que se han ajustado los modelos.

3.4. Análisis estadístico y predicción de las $PM_{2,5}$

Se ha seguido el mismo procedimiento que anteriormente.

Se representan en la Figura 30 los datos de la concentración a lo largo del tiempo con un color diferente para cada región.

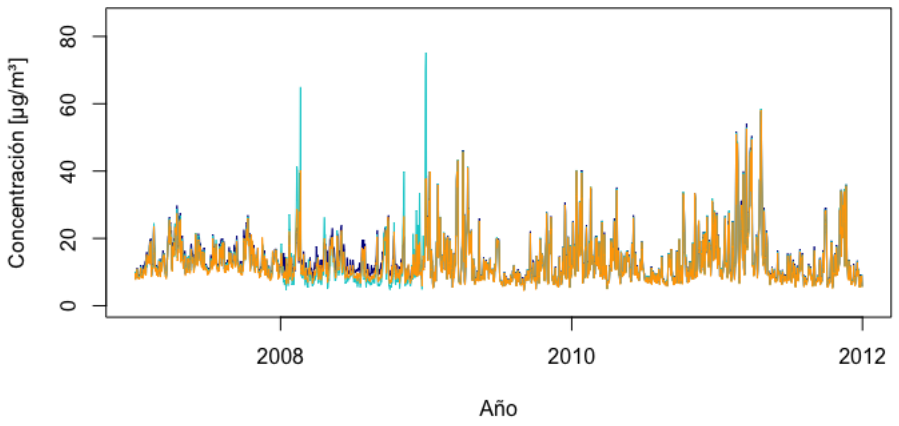


Figura 30: Concentración de PM_{10} en el periodo de estudio en tres regiones: Región 1 (azul oscuro), Región 2 (azul claro), Región 3 (naranja).

Al igual que en el análisis de las PM_{10}, de la observación de la Figura 24, es claro que en este conjunto de datos hay menor diferencia entre los valores de las tres regiones, siguiendo las tres la misma trayectoria a lo largo del periodo de estudio.

En la Tabla 9 se muestran los principales estadísticos de estudio.

Tabla 9: Tabla con estadísticos calculados para analizar la tendencia general de la serie.

año	media	mediana	rango intercuartílico	índice Yule - Kendall
2007	13.681	12.833	5.221	0.213
2008	12.240	10.568	5.251	0.324
2009	12.517	10.092	7.098	0.339
2010	13.404	11.530	7.292	0.244
2011	14.316	10.674	8.253	0.414

Los valores de las medianas son menores que los valores de las medias y distan de estas una cantidad que se ve incrementada a lo largo de los años. Esto nos indica que hay mayor irregularidad de los datos por encima de la mediana, y que esta irregularidad aumenta con el paso del tiempo. Esto se puede observar claramente en la Figura 30. Que el índice de Yule-Kendall sea positivo y bastante alejado de cero, también nos indica que hay una clara tendencia de los datos a estar por encima de la mediana.

Los valores de la media oscilan en torno a un valor de alrededor de $13\frac{\mu g}{m^3}$, teniendo la serie una tendencia creciente a partir de 2008.

En este caso, los valores de los rangos intercuartílicos presentados en la Tabla 9 son menores que los obtenidos en el análisis de las series temporales de los otros contaminantes. Esto indica una mayor constancia o pequeña oscilación de los valores en torno a un valor central.

Se representa en la Figura 31 un diagrama de cajas y bigotes de las medias mensuales de todos los años incluyendo los datos de las tres regiones.

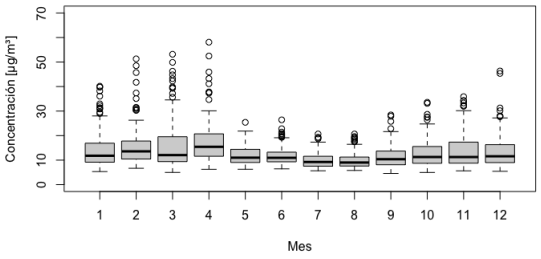
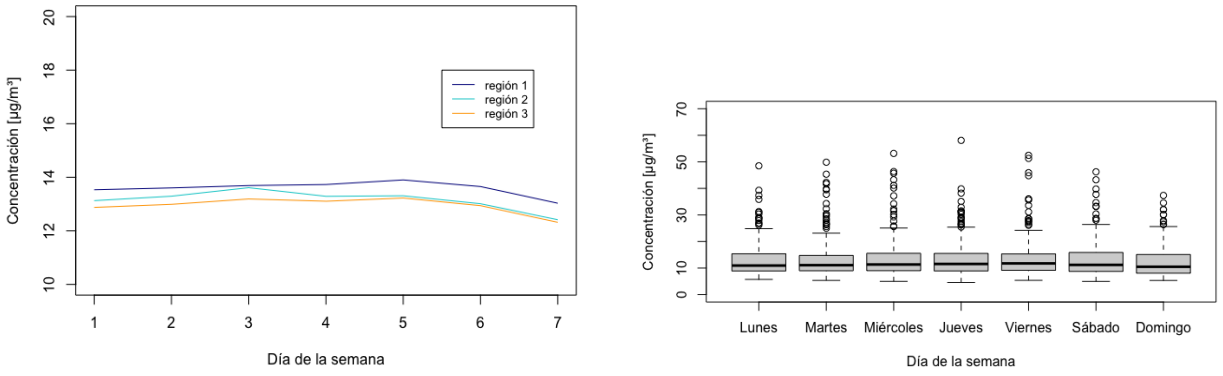


Figura 31: Diagrama de cajas y bigotes para representar la evolución de la serie a lo largo de los meses en un año. Se utilizan los valores de los cinco años por cada mes.

Este patrón anual de la concentración de $PM_{2,5}$ en la atmósfera de Londres recuerda al presentado en la Figura 25. Este patrón, al igual que los patrones de las concentraciones de NO_2 y de las PM_{10}, se puede explicar principalmente por dos factores: la actividad humana y la inversión térmica.

En la Figura 31 también se puede ver la gran dispersión de los valores por encima de la mediana, la pequeña oscilación de los valores en torno a un valor central y los valores bajos del rango intercuartílico.

Se representan los datos de media por semana y un diagrama de cajas y bigotes con todos los valores en función de los días de la semana.



(a) (b)

Figura 32: Tendencia semanal de la serie temporal. (a) Valores de la concentración media de $PM_{2,5}$ en función del día de la semana a lo largo del periodo de estudio. (b) Diagrama de cajas y bigotes con todos los valores a lo largo del periodo de estudio en función del día de la semana.

En la Figura 32 se ve cómo existe una tendencia a lo largo de los días de la semana. Durante los días laborables el valor de la concentración del $PM_{2,5}$ es constante mientras que el fin de semana decrece notablemente.

Este decrecimiento se debe en gran medida al descenso o cese de la actividad industrial y el uso de vehículos, que son dos de los grandes emisores del contaminante $PM_{2,5}$.

En resumen, del análisis estadístico se pueden extraer las siguientes conclusiones:

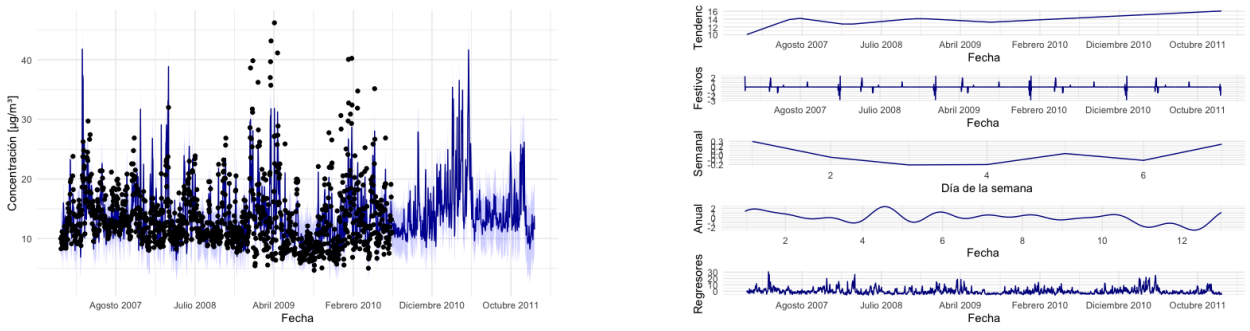
- El análisis de la concentración de las $PM_{2,5}$ es muy similar al de las PM_{10}, siendo valores más elevados los de la concentración de las PM_{10}. La tendencia de las concentraciones de las $PM_{2,5}$ es creciente a partir de 2008.
- Existe un ciclo anual; los valores en los meses de verano son menores que el valor medio y los valores en los meses de invierno son mayores que el valor medio.
- Existe un ciclo semanal; los valores durante los días laborables mantienen un valor constante y cuando llega el fin de semana hay un decrecimiento de concentración de $PM_{2,5}$ en la atmósfera.

3.4.1. Predicción con el algoritmo Prophet. Segundo acercamiento

De nuevo, sólo se presentan los resultados del segundo acercamiento porque son los que mejor calculan las componentes del modelo. El resto de predicciones y estadísticos se presentan en el Anexo D.

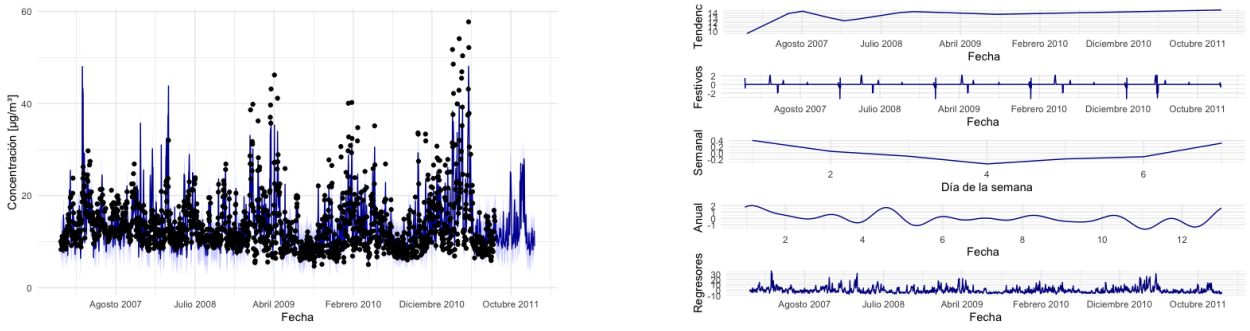
En este caso, se ha decidido utilizar regresores en el ajuste del modelo para observar el funcionamiento del algoritmo cuando se le añade información sobre otros factores relacionados con el contaminante principal a analizar, en este caso las $PM_{2,5}$. Se han añadido como regresores los valores de la región uno de las series temporales de concentración de NO_2, O_3 y PM_{10} a lo largo de todo el periodo de estudio.

Al añadir regresores, es importante dar valores para el periodo de tiempo con el que se ajusta el modelo, pero también para el periodo de tiempo sobre el que se van a hacer las predicciones.



(a) (b)

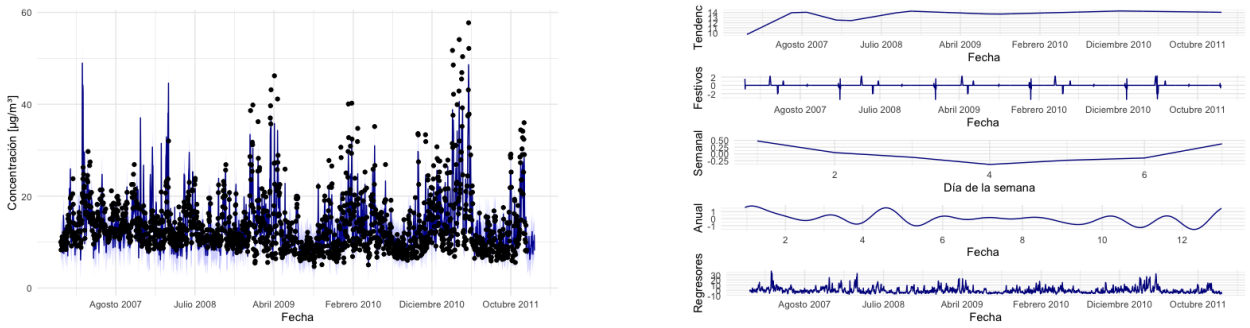
Figura 33: Modelo ajustado con los datos de la región 1 del $PM_{2,5}$ desde el 2007-01-01 hasta el 2010-07-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.



(a) (b)

Figura 34: Modelo ajustado con los datos de la región 1 del $PM_{2,5}$ desde el 2007-01-01 hasta el 2011-08-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

En esta ocasión las tres predicciones son bastante similares, lo que sugiere una regularidad de tendencia de la serie de datos con la que se ajustan los modelos.



(a) (b)

Figura 35: Modelo ajustado con los datos de la región 1 del $PM_{2,5}$ desde el 2007-01-01 hasta el 2011-12-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Valores predichos de la serie en azul oscuro, rango de error en azul claro y en forma de puntos negros los valores con los que se ajusta el modelo. (b) Componentes del modelo que se utiliza para la predicción.

Se puede observar que, en las tres figuras, la tendencia semanal indica un crecimiento de los valores durante el fin de semana, lo que no concuerda con la Figura 32, presentada en el análisis inicial que se ha realizado sobre la concentración de las $PM_{2,5}$ a lo largo del periodo de estudio.

Se muestra en la Tabla 10 los errores absolutos medios, que se utilizan para medir la bondad de los modelos de ajuste y predicción, en función del paso del proceso que se sigue en el segundo acercamiento.

Tabla 10: Error absoluto medio de los valores predichos por el algoritmo Prophet durante el acercamiento dos con respecto a los valores reales.

paso	EAM
1	6.418
2	6.664
3	6.703

En este caso, curiosamente, el error absoluto medio es mayor cuantos más valores se proporcionen para hacer el ajuste del modelo. Es posible que estemos en un caso de sobreajuste, en el que, por proporcionar demasiados parámetros, incluyendo los regresores, para ajustar el modelo, se prioriza una predicción exacta individual de cada uno de los valores a predecir, en lugar de un ajuste conjunto adecuado de las componentes.

4. Conclusiones

En este trabajo, se ha llevado a cabo un análisis exhaustivo de la concentración de cuatro contaminantes atmosféricos clave (NO_2, O_3, PM_{10} y $PM_{2,5}$) en el núcleo urbano de la ciudad de Londres. A través de un enfoque basado en estudiar la serie para posteriormente realizar predicciones mediante

el algoritmo Prophet de Facebook en el lenguaje de programación R, se ha logrado obtener una visión profunda de las tendencias y patrones en la concentración de estos contaminantes a lo largo del periodo de estudio. A continuación, se resumen las principales conclusiones obtenidas:

- Correlación entre las regiones: Figuras como la Figura 4 y tablas como la Tabla 1, han revelado una fuerte correlación positiva entre las concentraciones de NO_2, O_3, PM_{10} y $PM_{2,5}$ en las diferentes regiones de estudio. Dicho de otra forma, los datos en las tres regiones de estudio siguen la misma tendencia y trayectoria. Los patrones y las tendencias en los tres conjuntos de datos son muy parecidos y la variación en una región permite explicar en gran medida la variación en las otras regiones. Por lo tanto, se pueden analizar los movimientos generales de la serie observando únicamente los valores de una de las regiones, o alternativamente utilizando para cada día el valor medio de las tres regiones.

- Comparación entre contaminantes: Gracias a los estadísticos presentados en las Tablas 2, 5, 7 y 9 podemos concluir que los valores más altos son los de la concentración del NO_2, luego los del O_3, luego los de las PM_{10} y por último los de las $PM_{2,5}$. En todos los casos las trayectorias de la representación de los valores siguen un ciclo anual oscilando en torno a un valor central correspondiente a cada componente. En el caso de las PM, los valores del rango intercuartílico son menores que en los casos del NO_2 y del O_3. Esto quiere decir que la oscilación en torno al valor medio es de amplitud menor en los casos de las PM. Los valores de los índices de Yule-Kendall son cercanos a cero en los casos del NO_2 y del O_3, lo que quiere decir que los valores y los valores atípicos están equilibradamente distribuidos por encima y por debajo de la mediana. En el caso de las PM sin embargo el índice de Yule-Kendall es positivo, lo que indica que los valores tienen una mayor tendencia a estar por encima de la mediana. Esto indica que hay más valores atípicos por encima de la mediana. Esto se observa claramente en figuras como las Figuras 24 o 25.

- Ciclos anual y semanal: Todos los contaminantes salvo el ozono siguen un patrón en el que hay más concentración en los meses de invierno que en los meses de verano. También describen un patrón semanal con concentraciones constantes durante los días laborables y decrecimiento general de las concentraciones el fin de semana. Estos ciclos se pueden explicar mediante factores como las condiciones meteorológicas, la actividad humana o las inversiones térmicas.

- Particularidad del ozono: El análisis del ozono es inverso, siguiendo un patrón anual en el que la concentración es mayor durante los meses de abril, mayo y de verano y menor durante los meses de invierno y un patrón semanal con una subida de valores durante los fines de semana. Estos ciclos se pueden explicar mediante factores como la radiación solar, las condiciones meteorológicas o la reducción de la contaminación de óxidos de nitrógeno NO_x y compuestos orgánicos volátiles durante los meses de verano. A pesar de haber mayor radiación solar en los meses de verano, también hay menos contaminantes con los que interactuar para producir ozono. En los meses de abril y mayo la radiación solar aumenta y la producción de contaminantes se mantiene prácticamente constante porque son meses laborales y escolares, lo que supone un uso constante de transportes e industria.

- Algoritmo Prophet: Las predicciones más satisfactorias se consiguen cuantos más días consecutivos se utilicen para ajustar el modelo. Sin embargo, si se proporcionan demasiados datos se puede producir un sobreajuste, caso en el que algoritmo se preocupa prioritariamente de predecir los valores más cercanos sin calcular adecuadamente las componentes del modelo. Además, se ha observado que con la inclusión de la serie temporal conocida como único argumento del algoritmo

es suficiente para obtener predicciones precisas. Las conclusiones de los acercamientos de estudio que se han realizado con el algoritmo Prophet son las siguientes:

- Si se ajusta el modelo con pocos valores muy separados en el tiempo las componentes del modelo no son visualmente correctas. No son suficientes datos para que el modelo identifique una periodicidad anual ni semanal específica. Esto se puede observar en la Figura 39.
- Si se ajusta el modelo con pocos valores, la componente de los días festivos y eventos marca picos de influencia en el cambio de año. Hay un cambio de tendencia cada año y el algoritmo identifica las vacaciones de Navidad como factor influyente en este cambio de tendencia. Esto también se puede observar en la Figura 39.
- En ocasiones las componentes son completamente erróneas o se alejan visualmente bastante de la realidad. Esto se ha asociado a una mala selección de los días aleatorios con los que se ajusta el modelo. Esto por ejemplo pasa en la Figura 40.
- En general, cuantos más datos utilicemos para ajustar el modelo, más se acercan los resultados de las componentes a la realidad.
- La componente de la tendencia general de la serie no está bien definida en ningún caso, no es suficiente el número de valores empleado para ajustar de manera correcta el modelo.
- Los valores específicos predichos siguen bien la trayectoria de la serie original porque se predicen los valores con los que se ajusta el modelo y los de días consecutivos, por lo cual la predicción es similar al valor real. Esto se ve en la fila de arriba de cada una de las gráficas.
- Las componentes de los modelos ajustados en los pasos uno y tres son muy similares, si no idénticas entre sí. Esto ocurre porque en el paso tres se ajusta el modelo con los valores predichos en el paso uno. La componente que más cambia entre los pasos uno y tres es la tendencia general de la serie, porque en el paso tres se utilizan más datos para ajustar el modelo que en el paso uno.
- En general los mejores ajustes son los del segundo acercamiento y de entre los del primer acercamiento los mejores son los del segundo paso.

En resumen, este estudio proporciona una comprensión profunda de la evolución de la concentración de los contaminantes atmosféricos en el entorno urbano de Londres a lo largo del periodo de estudio, así como los patrones que la serie puede seguir en los años directamente consecutivos a los del estudio. Estos procedimientos son utilizables para las tomas de decisiones relacionadas con la gestión de calidad del aire y la implementación de estrategias de reducción de la contaminación en áreas urbanas. El enfoque utilizado, que combina análisis estadísticos y herramientas computacionales de predicción de datos, puede ser aplicado a otras ciudades para evaluar y abordar problemas similares de calidad del aire.

5. Referencias

BIRINCI, Enes; ÖZDEMIR, Emrah Tuncay y DENIZ, Ali, 2023. An investigation of the effects of sand and dust storms in the North East Sahara Desert on Turkish airports and PM10 values: 7 and 8 April, 2013 events. *Environmental Monitoring and Assessment*. **195**, n.º 6. Disp. desde DOI: 10.1007/s10661-023-11288-5.

DONNELLY, Aoife; MISSTEAR, Bruce y BRODERICK, Brian, 2011. Application of nonparametric regression methods to study the relationship between NO2 concentrations and local wind direction and speed at background sites. *Science of the Total Environment*. **409**, n.º 6, 1134-1144. Disp. desde DOI: 10.1016/j.scitotenv.2010.12.001.

FERNÁNDEZ-DUQUE, Beatriz; PÉREZ, Isidro A.; GARCÍA, M. Ángeles; PARDO, Nuria y SÁNCHEZ, M. Luisa, 2019. Annual and seasonal cycles of CO2 and CH4 in a Mediterranean Spanish environment using different kernel functions. *Stochastic Environmental Research and Risk Assessment*. **33**, n.º 3, 915-930. Disp. desde DOI: 10.1007/s00477-019-01655-5.

GARCÍA, M.; RAMÍREZ, H.; ULLOA, H.; ARIAS, S. y PÉREZ, A., 2012. Las inversiones térmicas y la contaminación atmosférica en la zona metropolitana de Guadalajara (México). *Instituto de Astronomía y Meteorología. Universidad de Guadalajara (México). Universidad de Santiago de Compostela (España)*.

GROLEMUND, Garrett y WICKHAM, Hadley, 2011. Dates and Times Made Easy with lubridate. *Journal of Statistical Software*. **40**, n.º 3, 1-25. Disponible también desde: https://www.jstatsoft.org/v40/i03/.

GUO, Ruyue; SHI, Guangming; ZHANG, Dan; CHEN, Yang; PENG, Chao; ZHAI, Chongzhi y YANG, Fumo, 2024. An observed nocturnal ozone transport event in the Sichuan Basin, Southwestern China. *Journal of Environmental Sciences (China)*. **138**, 10-18. Disp. desde DOI: 10.1016/j.jes.2023.02.054.

KHOLODOV, Aleksei; KIRICHENKO, Konstantin; VAKHNIUK, Igor; FATKULIN, Anvir; TRETYAKOVA, Maria; ALEKSEIKO, Leonid; PETUKHOV, Valeriy y GOLOKHVAST, Kirill, 2022. Measurement of PM2.5 and PM10 Concentrations in Nakhodka City with a Network of Automatic Monitoring Stations. *Aerosol and Air Quality Research*. **22**, n.º 10. Disp. desde DOI: 10.4209/aaqr.220040.

LIU, Zhiqiang; HU, Kun; ZHANG, Kun; ZHU, Shengnan; WANG, Ming y LI, Li, 2023. VOCs sources and roles in O3 formation in the central Yangtze River Delta region of China. *Atmospheric Environment*. **302**. Disp. desde DOI: 10.1016/j.atmosenv.2023.119755.

MILFORD, Celia; TORRES, Carlos; VILCHES, Jon; GOSSMAN, Ann-Kathrin; WEIS, Frederik; SUÁREZ-MOLINA, David; GARCÍA, Omaira E.; PRATS, Natalia; BARRETO, África; GARCÍA, Rosa D.; BUSTOS, Juan J.; MARRERO, Carlos L.; RAMOS, Ramón; CHINEA, Nayra; BOULESTEIX, Thomas; TAQUET, Noémie; RODRÍGUEZ, Sergio; LÓPEZ-DARIAS, Jessica; SICARD, Michaël; CÓRDOBA-JABONERO, Carmen y CUEVAS, Emilio, 2023. Impact of the 2021 La Palma volcanic eruption on air quality: Insights from a multidisciplinary approach. *Science of the Total Environment*. **869**. Disp. desde DOI: 10.1016/j.scitotenv.2023.161652.

PALEI, Soumyak; DAS, Sreetama; SARKAR, Rajasree; CHAUDHURI, Amrita; DUTTA, Subhankar y NAYEK, Sumanta, 2023. Assessment of Air Quality with Respect to Particulate Matter (PM10, PM2.5) in Mining Industrial Areas of Keonjhar District, Odisha, and Its Public Health Implications. *Lecture Notes in Civil Engineering*. **323 LNCE**, 87-96. Disp. desde DOI: 10.1007/978-981-99-0823-3_9.

R CORE TEAM, 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Disponible también desde: https://www.R-project.org/.

ROY, Chaitri; RAVISHANKARA, A.R.; NEWMAN, Paul A.; DAVID, Liji M.; FADNAVIS, Suvarna; RATHOD, Sagar D.; LAIT, Leslie; KRISHNAN, R.; CLARK, Hannah y SAUVAGE, Bastien, 2023. Estimation of Stratospheric Intrusions During Indian Cyclones. *Journal of Geophysical Research: Atmospheres*. **128**, n.º 3. Disp. desde DOI: 10.1029/2022JD037519.

SICARD, Pierre; AGATHOKLEOUS, Evgenios; ANENBERG, Susan C.; DE MARCO, Alessandra; PAOLETTI, Elena y CALATAYUD, Vicent, 2023. Trends in urban air pollution over the last two decades: A global perspective. *Science of the Total Environment*. **858**. Disp. desde DOI: 10.1016/j.scitotenv.2022.160064.

TAYLOR, Sean y LETHAM, Benjamin, 2017. Forecasting at scale. *PeerJ Preprints*.

TAYLOR, Sean y LETHAM, Benjamin, 2021. *prophet: Automatic Forecasting Procedure*. Disponible también desde: https://CRAN.R-project.org/package=prophet. R package version 1.0.

WICKHAM, Hadley, 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. Disponible también desde: https://ggplot2.tidyverse.org.

WICKHAM, Hadley; FRANÇOIS, Romain; HENRY, Lionel; MÜLLER, Kirill y VAUGHAN, Davis, 2023. *dplyr: A Grammar of Data Manipulation*. Disponible también desde: https://CRAN.R-project.org/package=dplyr. R package version 1.1.2.

WUERTZ, Diethelm; SETZ, Tobias; CHALABI, Yohan y BOSHNAKOV, Georgi N., 2023. *timeDate: Rmetrics - Chronological and Calendar Objects*. Disponible también desde: https://CRAN.R-project.org/package=timeDate. R package version 4022.108.

Anexo A. Gráficas y estadísticos de las predicciones del algoritmo Prophet para los datos pertenecientes al NO_2.

Se presentan las gráficas por orden de año y número de datos utilizados para ajustar los modelos.

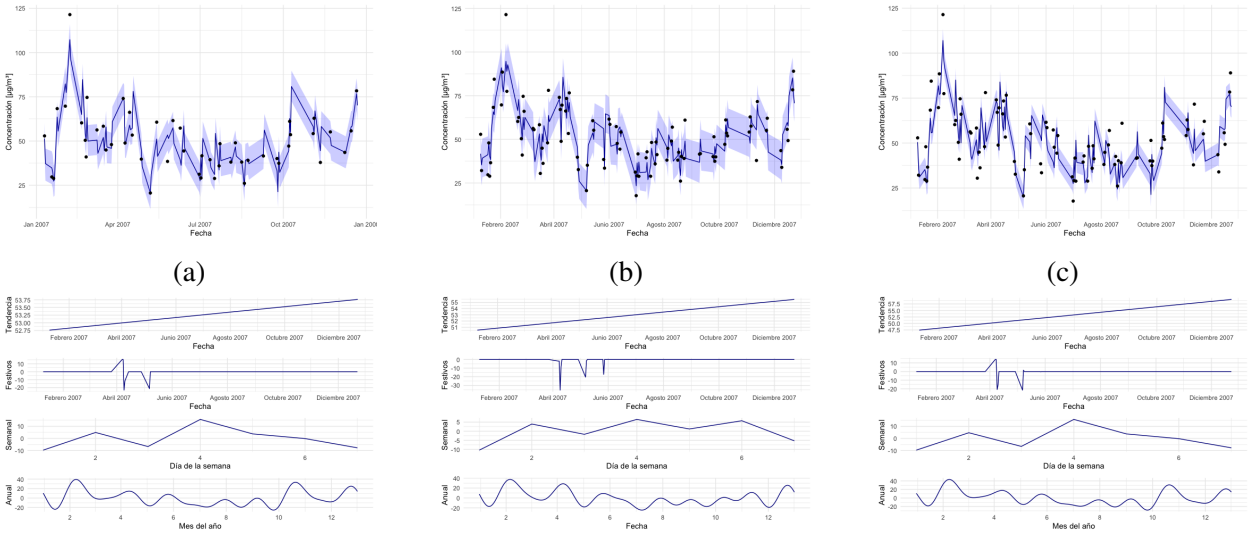
Por cada predicción se presentan dos gráficas:

1. La gráfica de la predicción en la que se muestran los valores predichos en azul oscuro con el rango de error en azul claro y los valores utilizados para ajustar el modelo utilizado para hacer la predicción con puntos negros.

2. La gráfica de las componentes de dicho modelo ajustado, en la que aparecen a su vez cuatro gráficas, en orden de arriba a abajo: Tendencia, efecto de los días festivos, periodicidad semanal, periodicidad anual.

Primer acercamiento

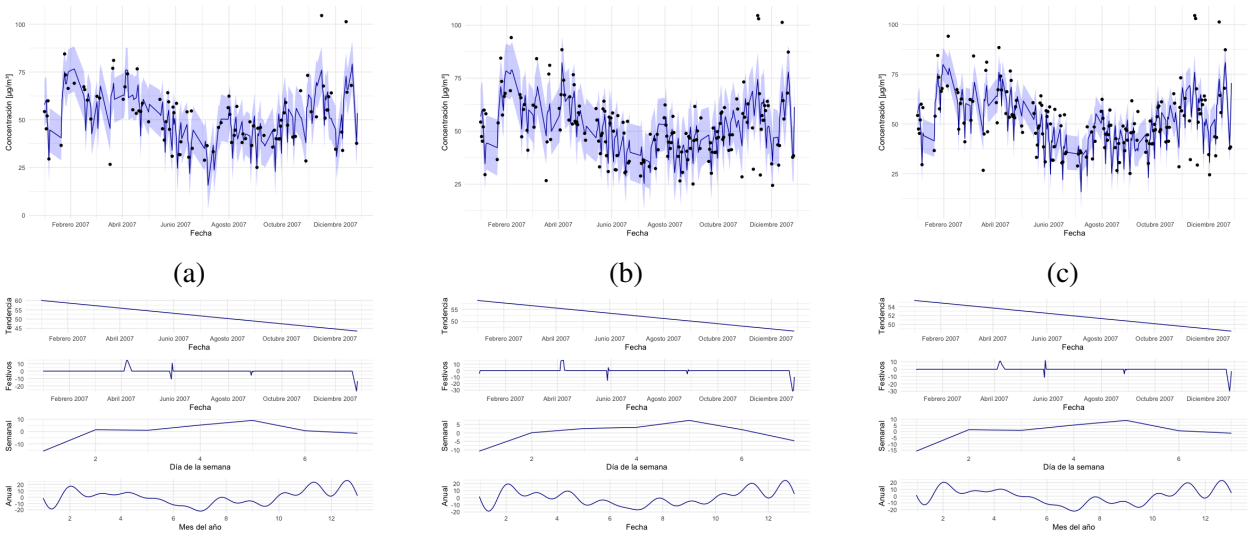
Fechas aleatorizadas de entre las fechas del año 2007.



(a) (b) (c)

(d) (e) (f)

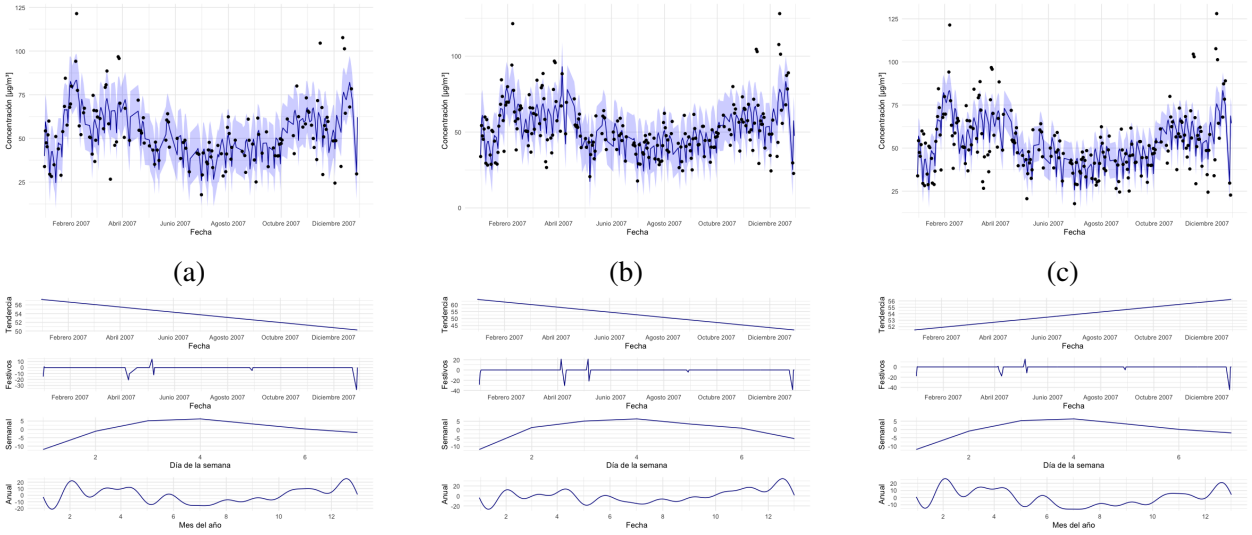
Figura 36: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 37: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

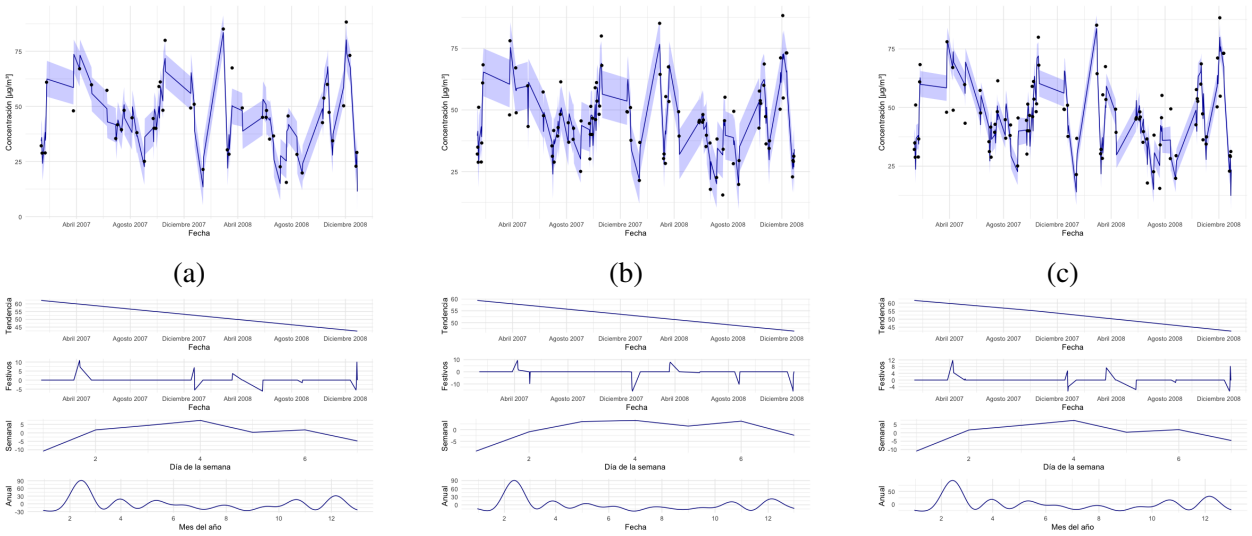


(a) (b) (c)

(d) (e) (f)

Figura 38: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

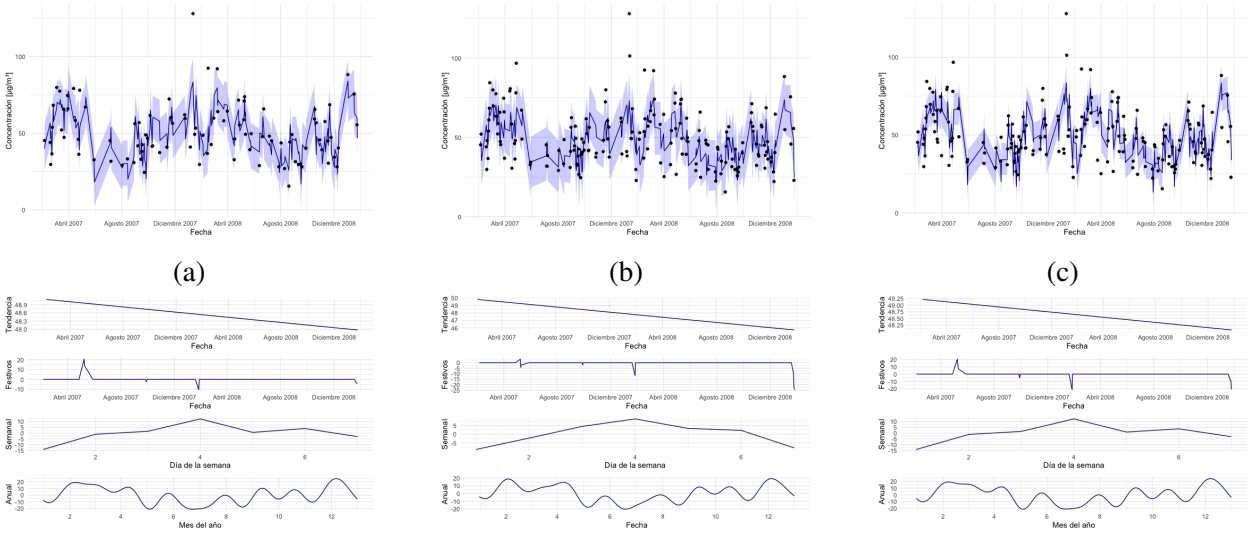
Fechas aleatorizadas de entre las fechas de los años 2007 y 2008.



(a) (b) (c)

(d) (e) (f)

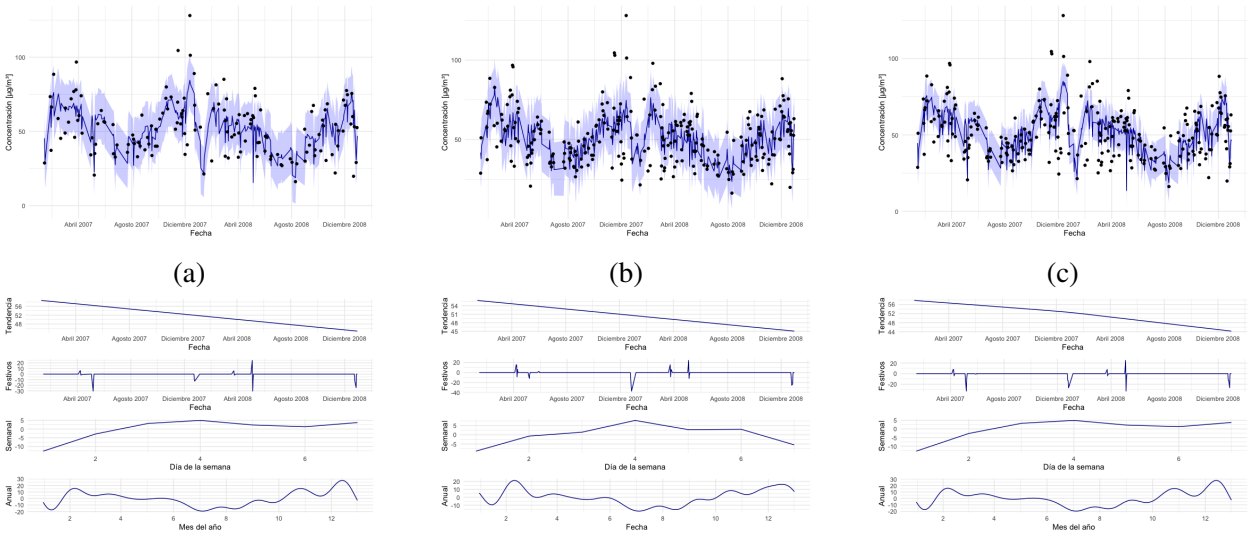
Figura 39: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

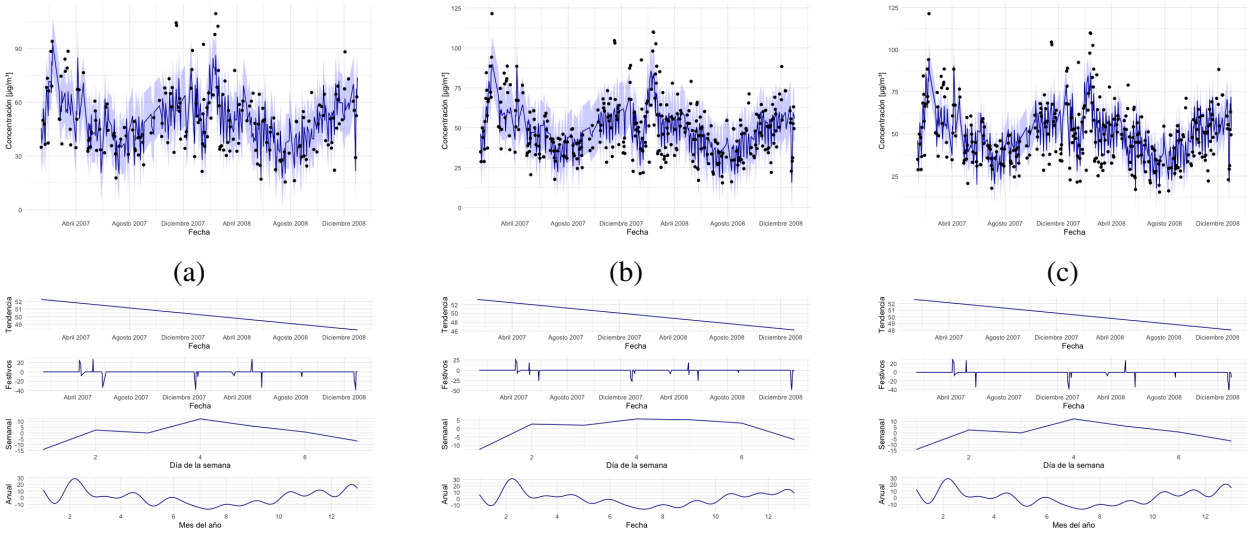
Figura 40: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 41: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

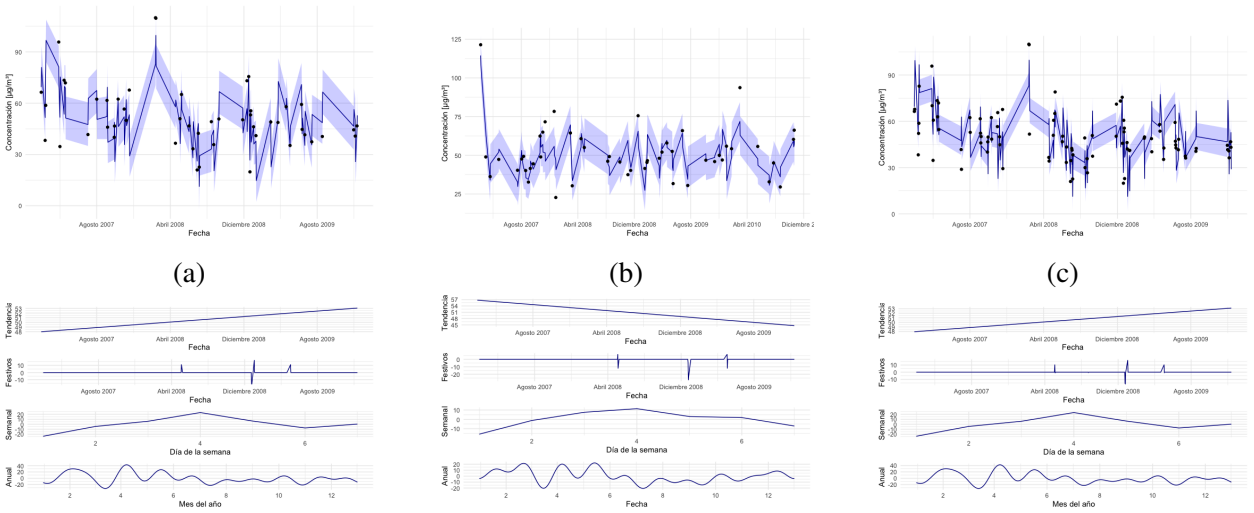


(a) (b) (c)

(d) (e) (f)

Figura 42: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. De izquierda a derecha paso 1, paso 2 y paso 3. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

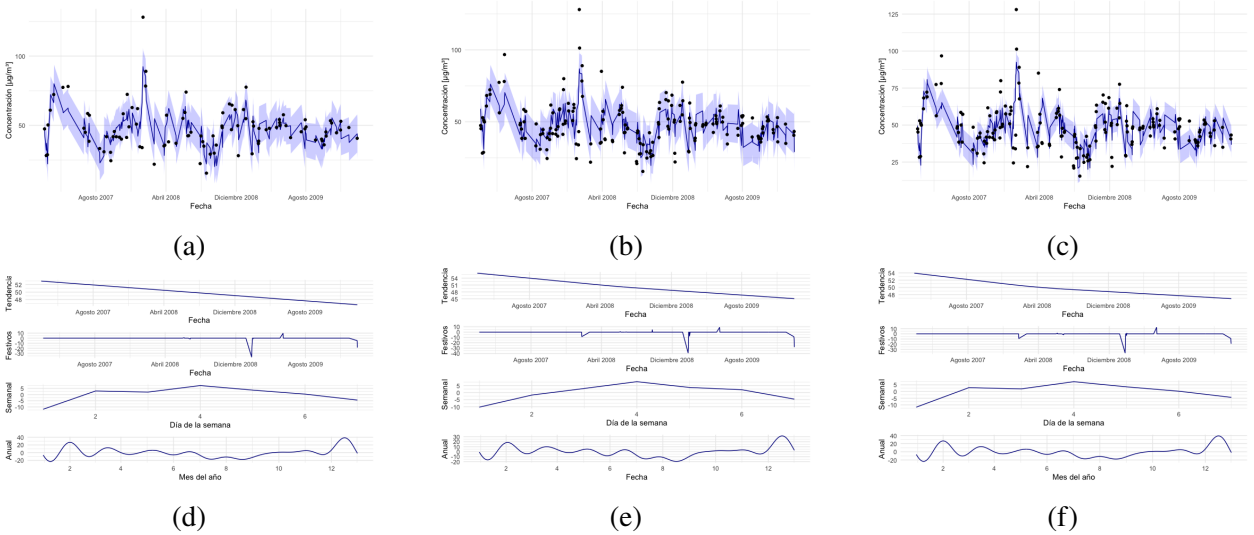
Fechas aleatorizadas de entre las fechas de los años 2007, 2008 y 2009.



(a) (b) (c)

(d) (e) (f)

Figura 43: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 44: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

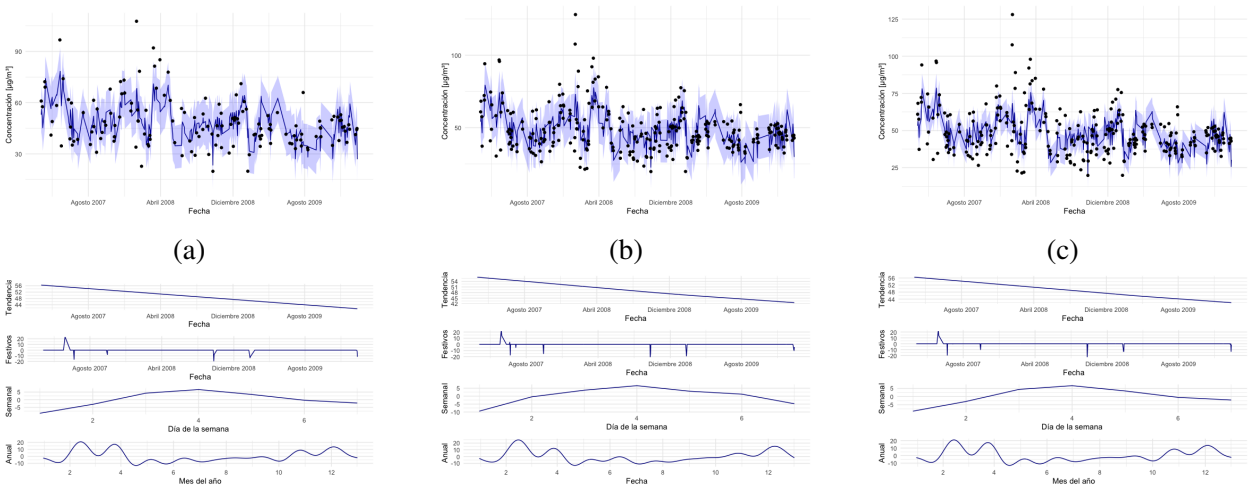


Figura 45: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

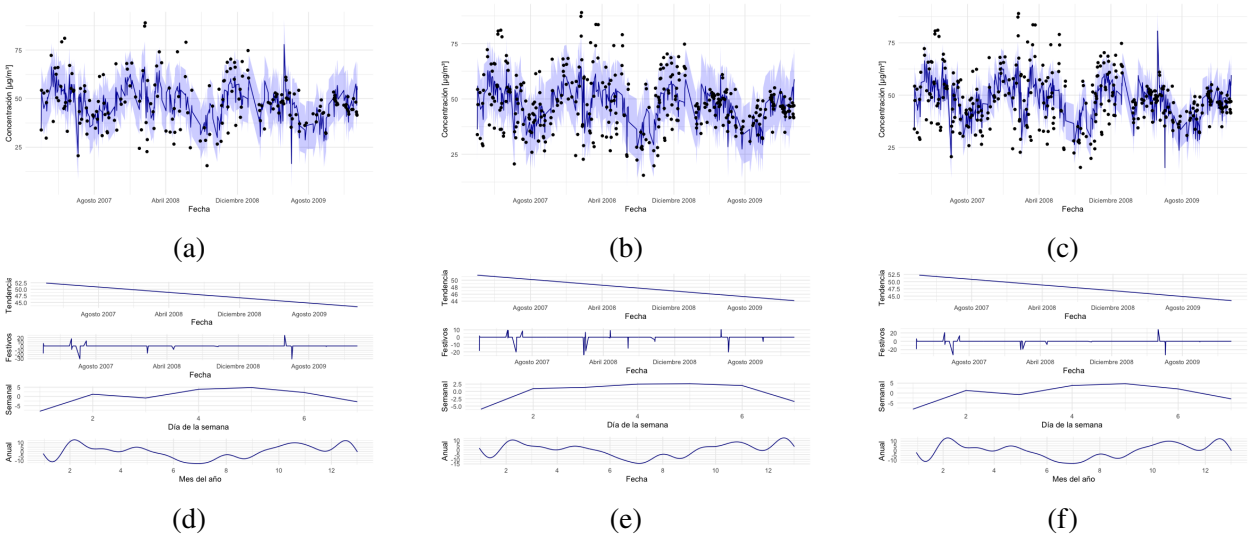
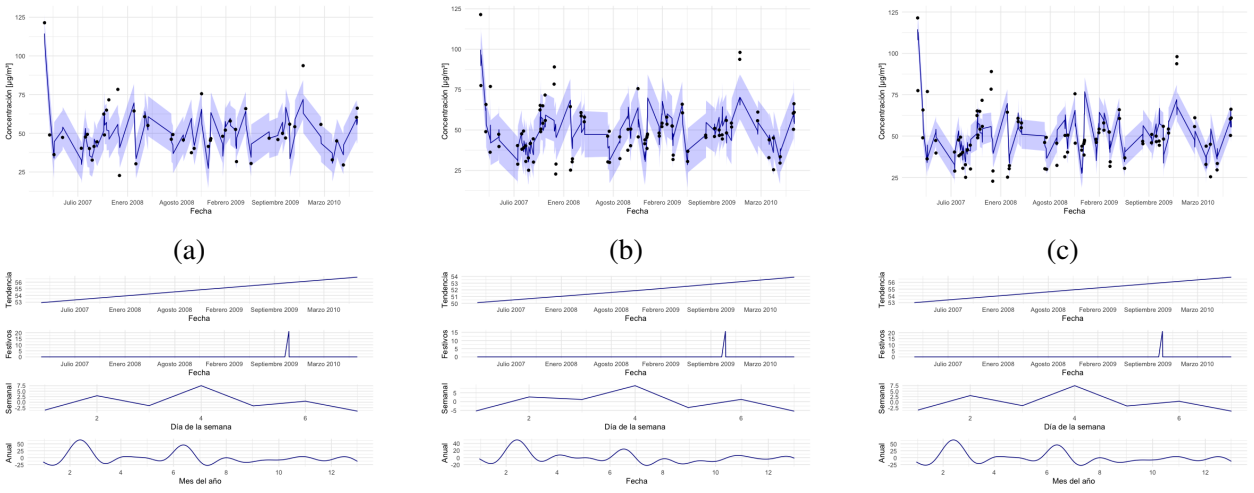


Figura 46: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

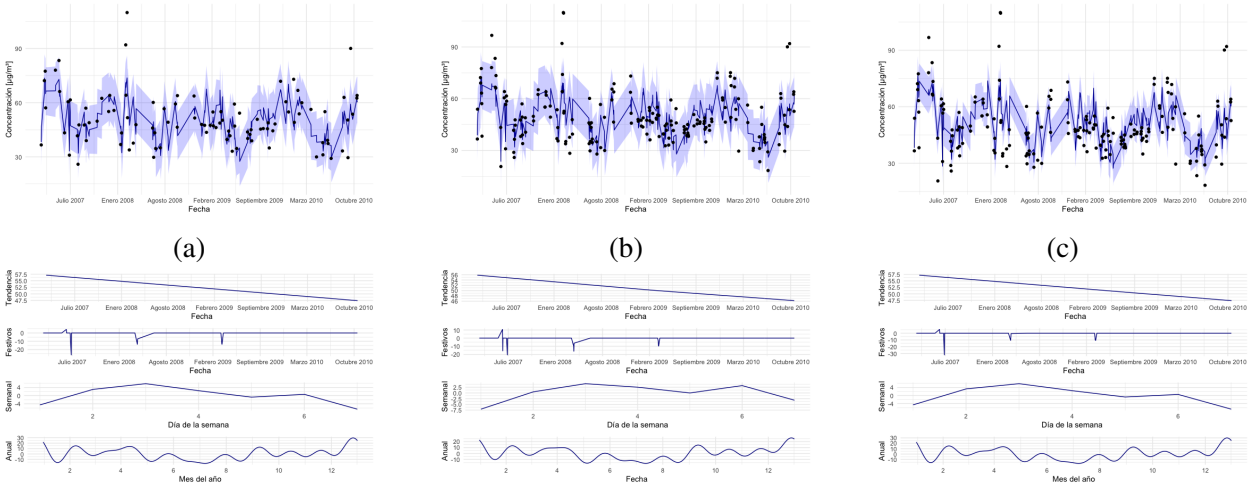
Fechas aleatorizadas de entre las fechas de los años 2007, 2008, 2009 y 2010.



(a) (b) (c)

(d) (e) (f)

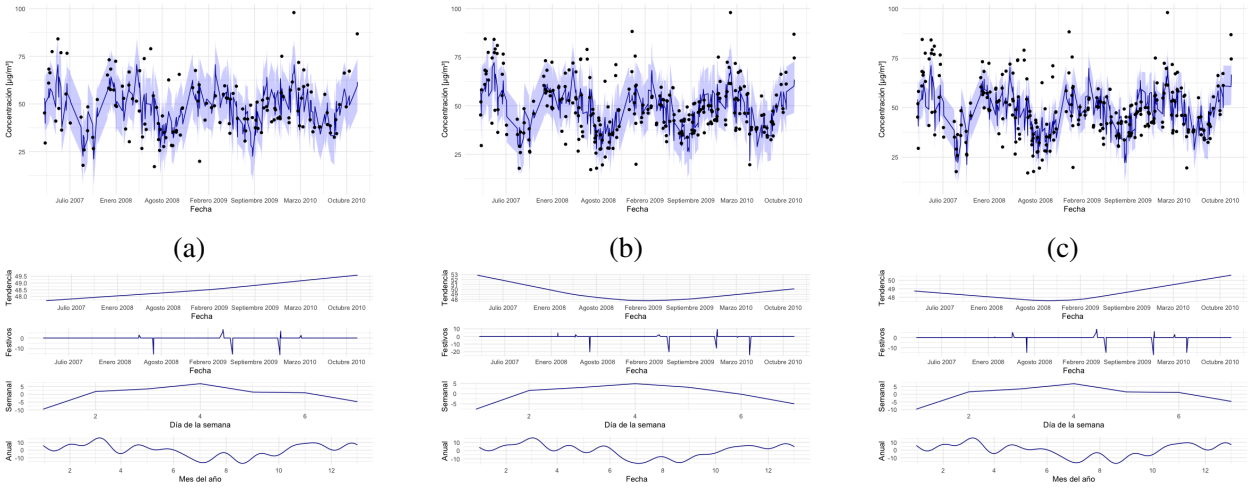
Figura 47: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

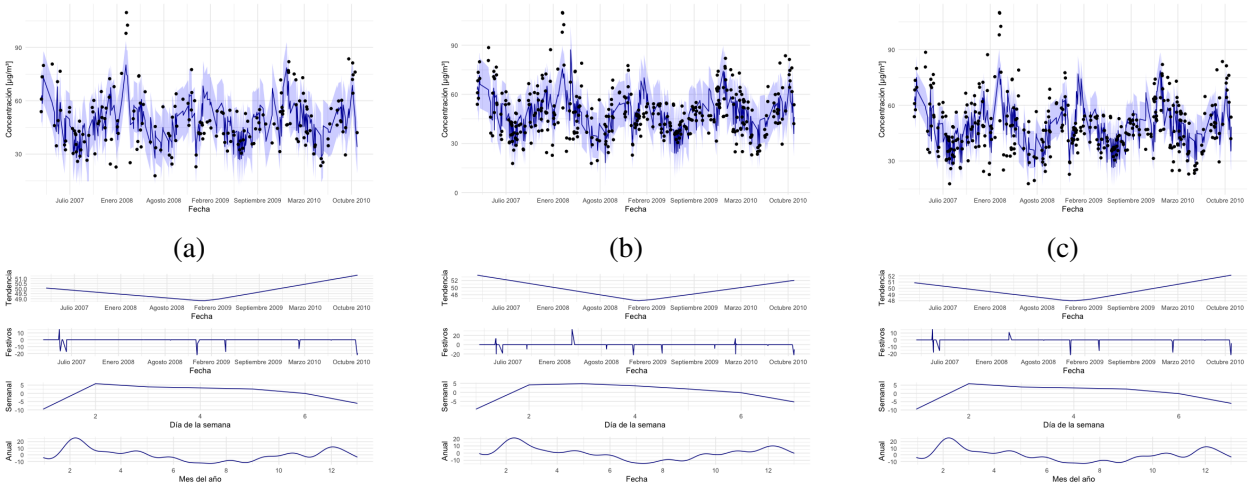
Figura 48: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 49: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

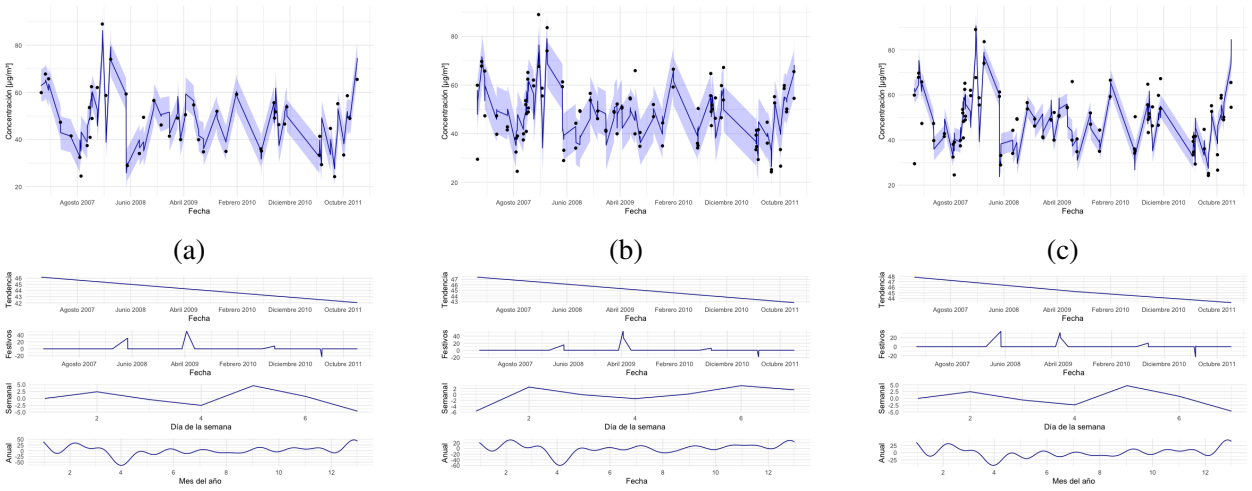


(a) (b) (c)

(d) (e) (f)

Figura 50: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

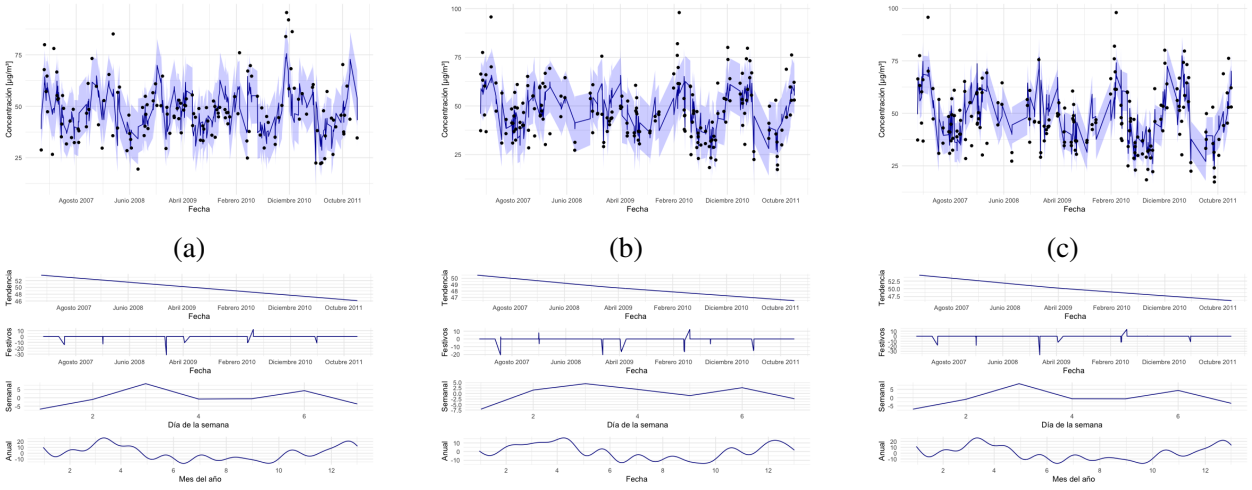
Fechas aleatorizadas de entre las fechas de los años 2007, 2008, 2009, 2010 y 2011.



(a) (b) (c)

(d) (e) (f)

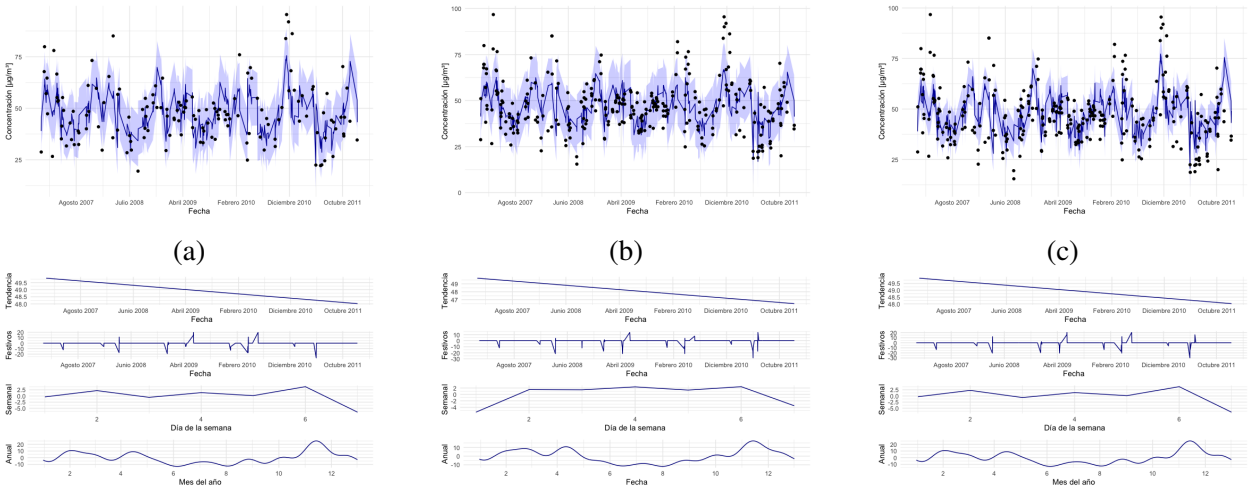
Figura 51: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

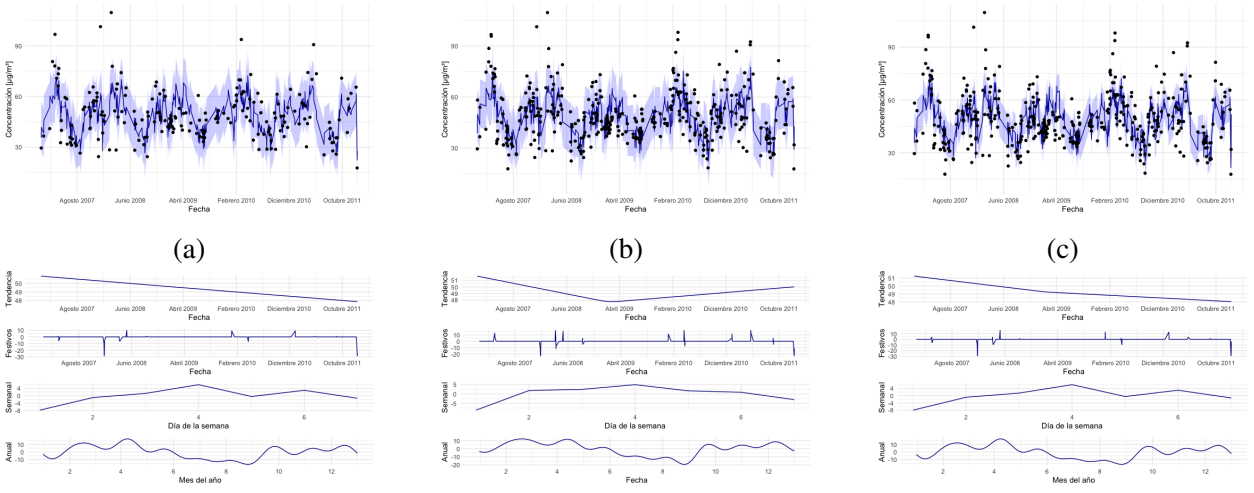
Figura 52: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 53: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

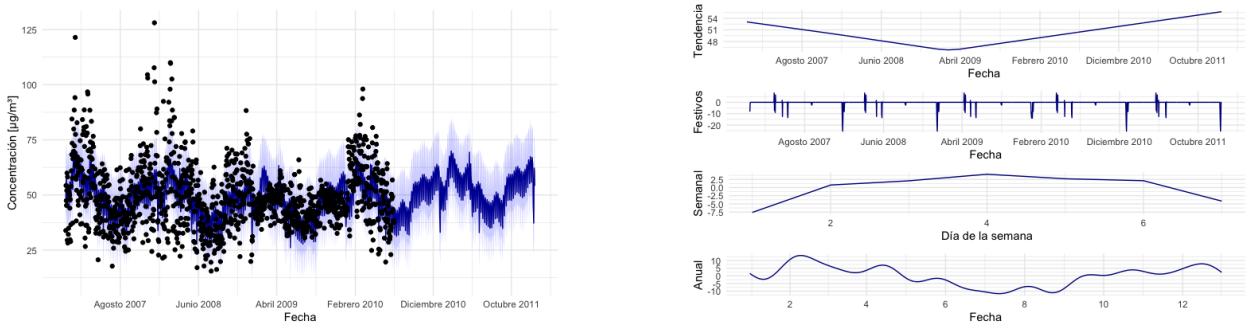


(a) (b) (c)

(d) (e) (f)

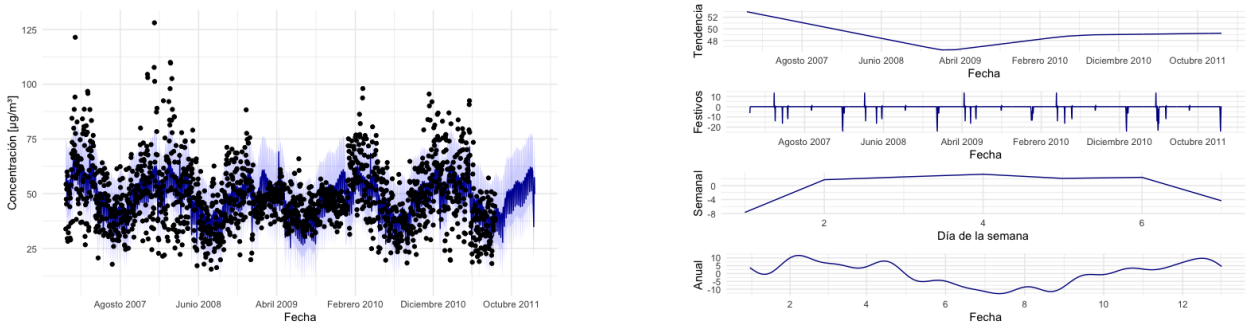
Figura 54: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

Segundo acercamiento



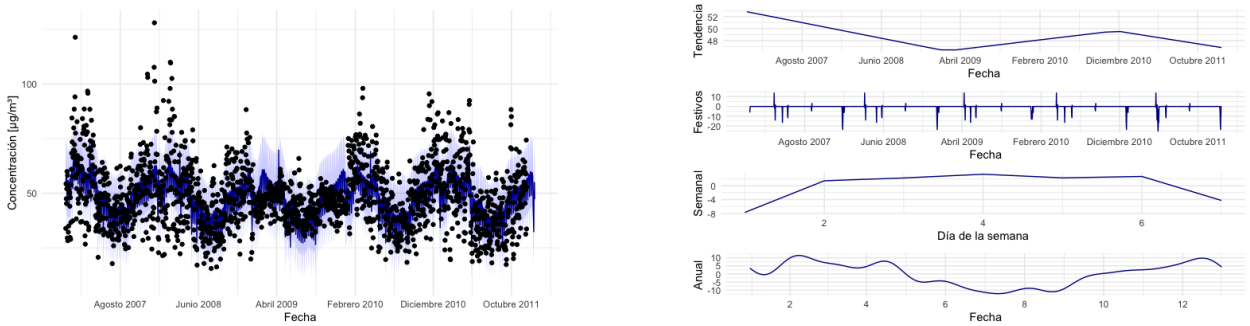
(a) (b)

Figura 55: Predicción del modelo ajustado con los datos de la región 1 del NO_2 desde el 2007-01-01 hasta el 2010-07-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.



(a) (b)

Figura 56: Predicción del modelo ajustado con los datos de la región 1 del NO_2 desde el 2007-01-01 hasta el 2011-08-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.



(a) (b)

Figura 57: Predicción del modelo ajustado con los datos de la región 1 del NO_2 desde el 2007-01-01 hasta el 2011-12-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.

Tabla 11: Errores absolutos medios de las predicciones realizadas utilizando los modelos ajustados. Paso representa el paso en el que se realiza la predicción. El año representa hasta qué año se incluye, partiendo desde 2007, para realizar la selección de días aleatorizados.

		Número de días aleatorizados			
	Paso	50	100	150	200
2007	1	8,737	8,778	9,691	
	2	8,200	8,708	9,516	
	3	10,123	9,694	9,884	
2008	1	6,802	10,539	11,579	10,631
	2	8,185	10,090	10,226	10,129
	3	8,819	10,827	11,624	10,881
2009	1	10,764	8,226	9,207	8,314
	2	8,062	8,415	9,316	8,435
	3	11,512	8,945	9,881	8,856
2010	1	9,097	9,215	8,594	9,154
	2	9,403	8,745	8,362	9,016
	3	10,446	9,477	8,981	9,378
2011	1	6,695	8,848	9,365	9,200
	2	6,816	8,247	8,988	8,959
	3	7,721	9,089	9,674	9,400

Anexo B. Gráficas y estadísticos de las predicciones del algoritmo Prophet para los datos pertenecientes al O_3.

Se presentan las gráficas por orden de año y número de datos utilizados para ajustar los modelos.

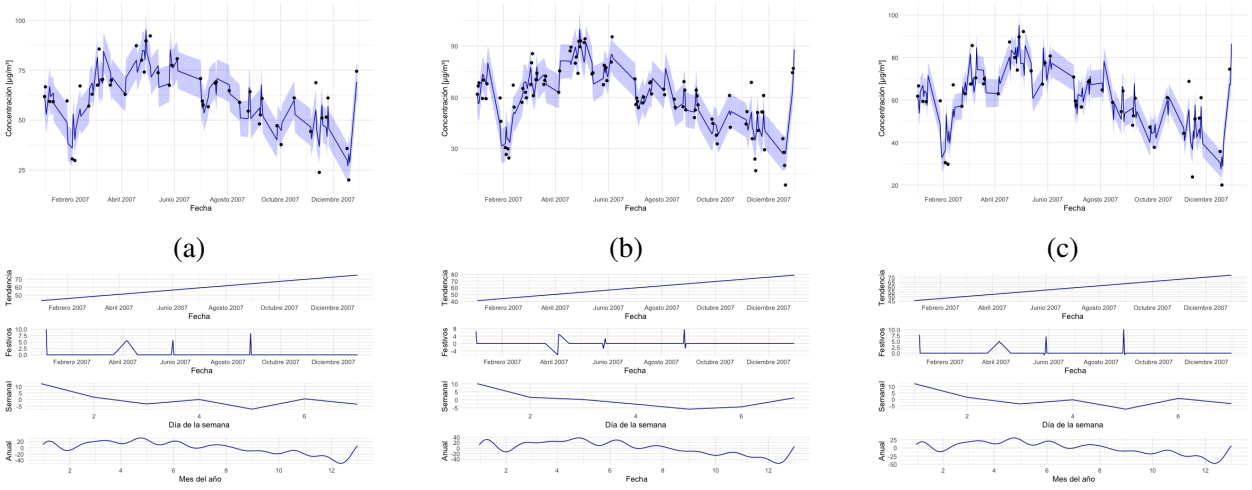
Por cada predicción se presentan dos gráficas:

1. La gráfica de la predicción en la que se muestran los valores predichos en azul oscuro con el rango de error en azul claro y los valores utilizados para ajustar el modelo utilizado para hacer la predicción con puntos negros.

2. La gráfica de las componentes de dicho modelo ajustado, en la que aparecen a su vez cuatro gráficas, en orden de arriba a abajo: Tendencia, efecto de los días festivos, periodicidad semanal, periodicidad anual.

Primer acercamiento

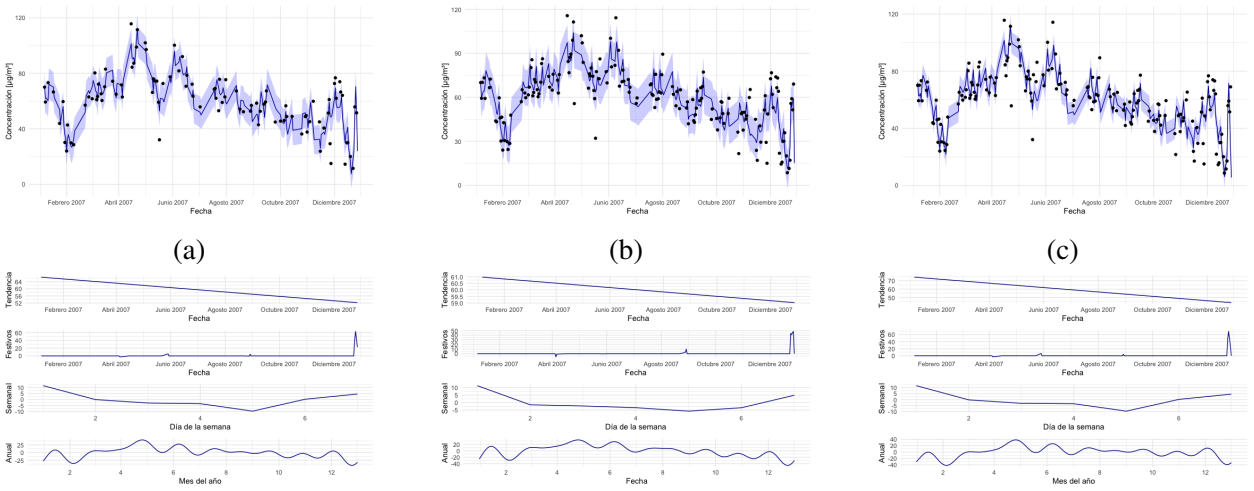
Fechas aleatorizadas de entre las fechas del año 2007.



(a) (b) (c)

(d) (e) (f)

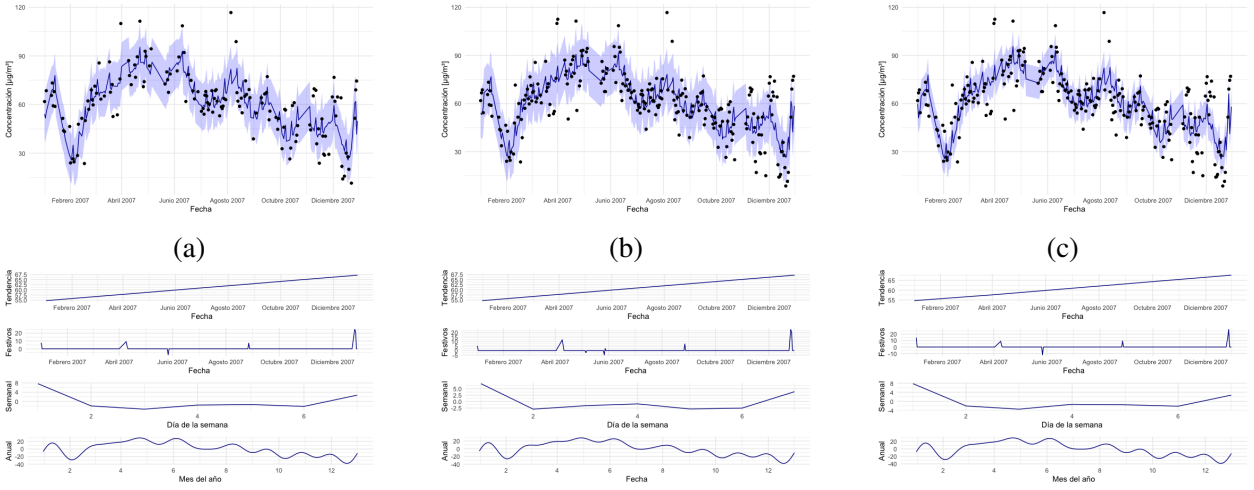
Figura 58: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 59: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

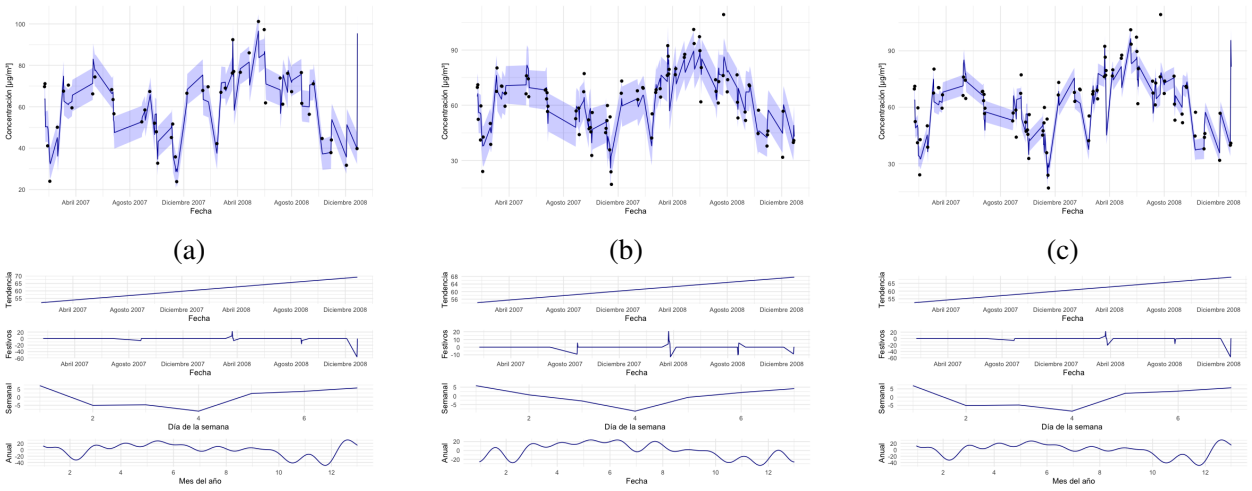


(a) (b) (c)

(d) (e) (f)

Figura 60: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

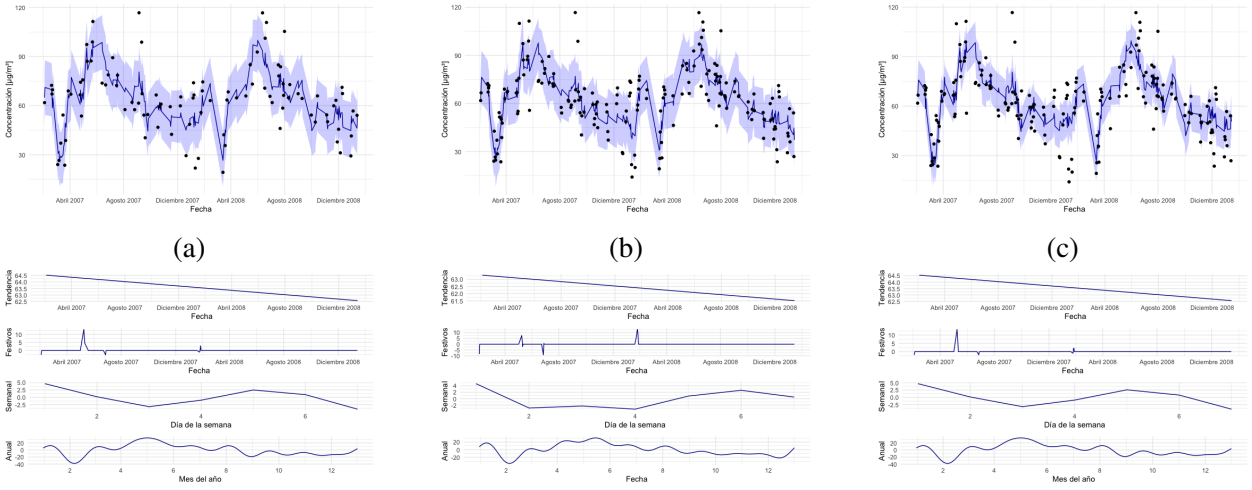
Fechas aleatorizadas de entre las fechas de los años 2007 y 2008.



(a) (b) (c)

(d) (e) (f)

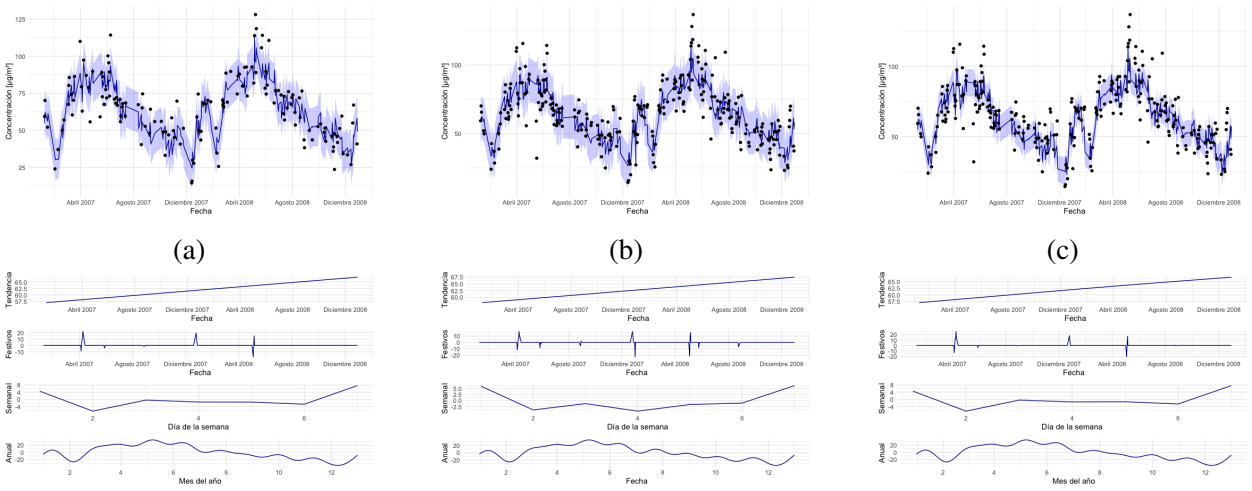
Figura 61: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

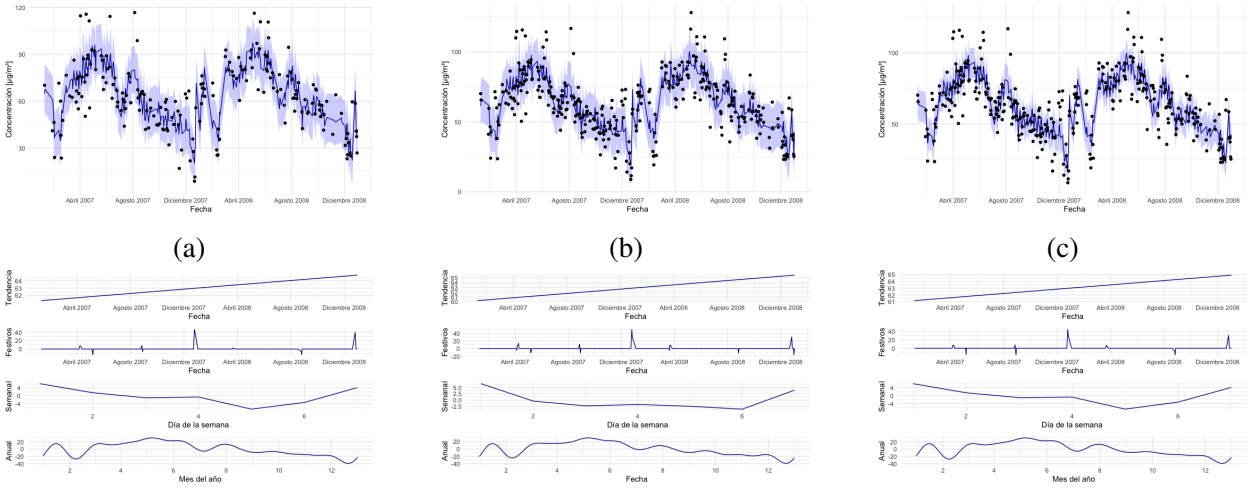
Figura 62: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 63: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 64: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. De izquierda a derecha paso 1, paso 2 y paso 3. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

Fechas aleatorizadas de entre las fechas de los años 2007, 2008 y 2009.

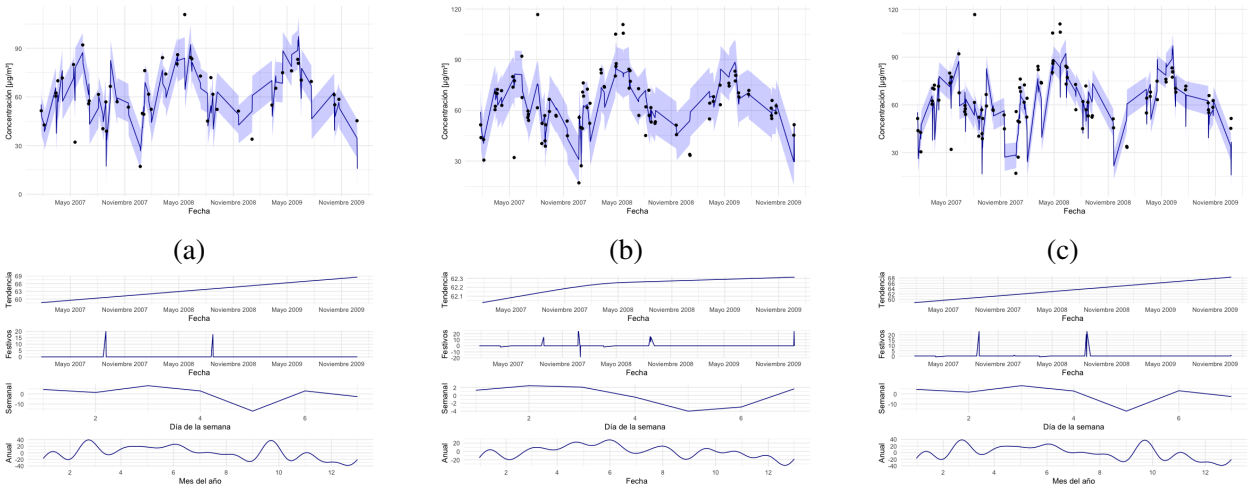


Figura 65: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

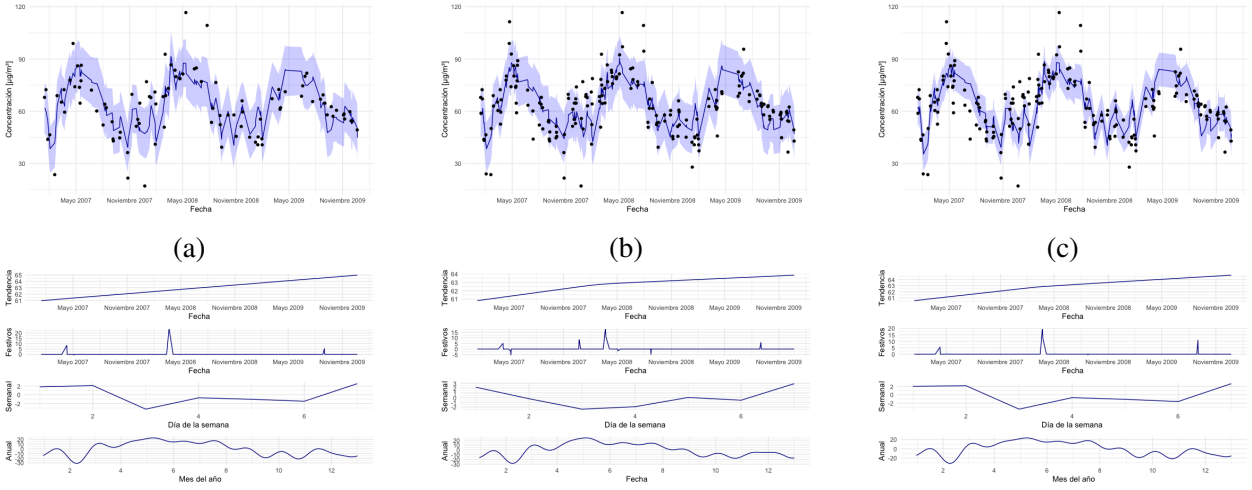
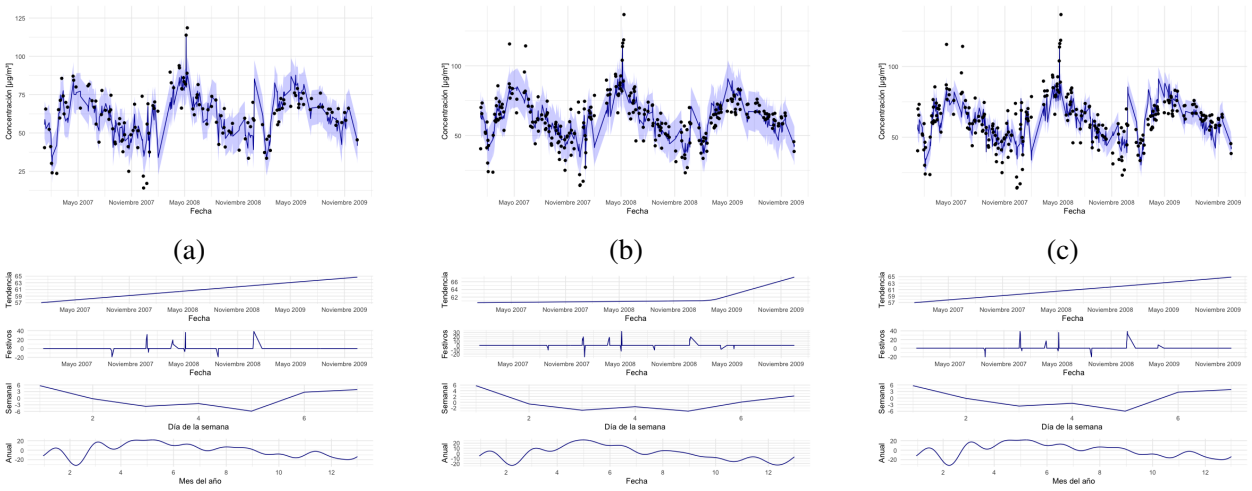


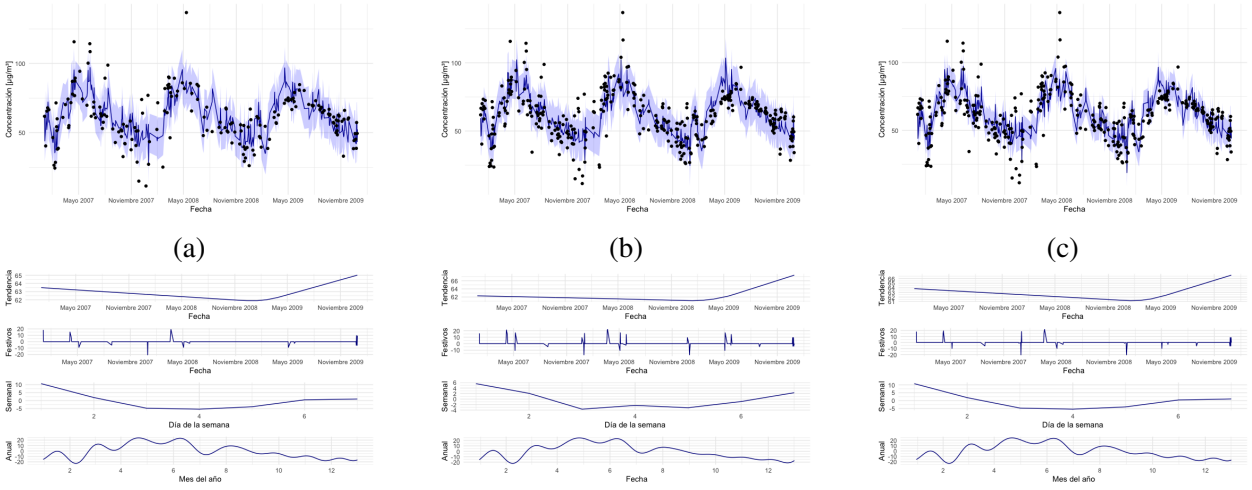
Figura 66: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 67: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

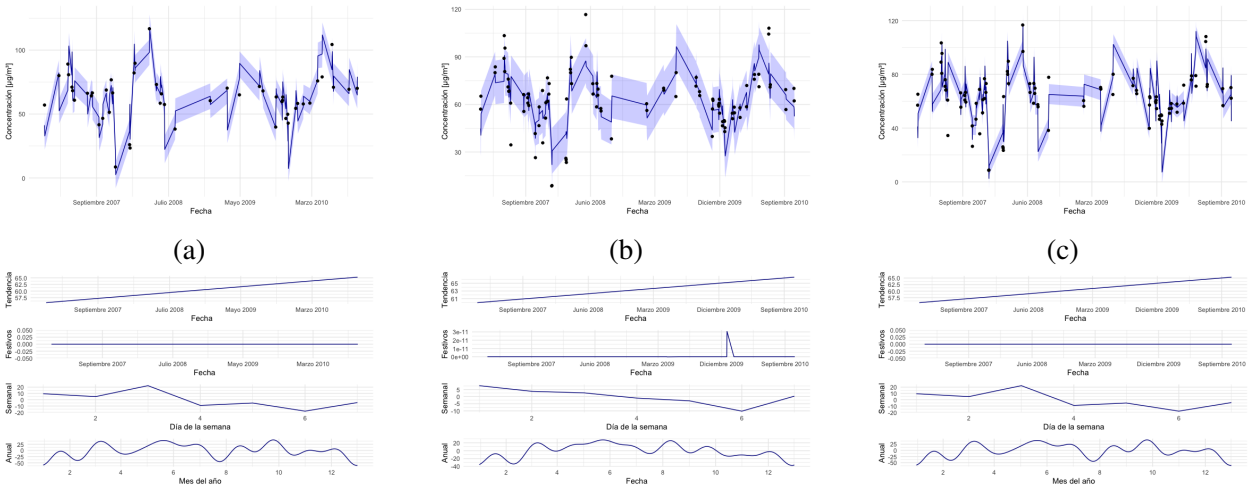


(a) (b) (c)

(d) (e) (f)

Figura 68: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

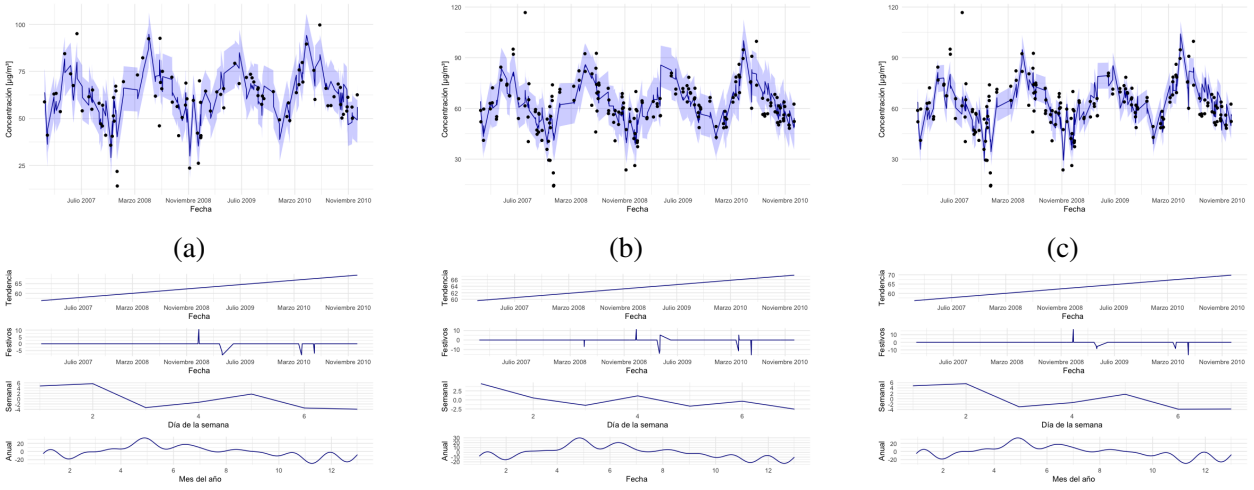
Fechas aleatorizadas de entre las fechas de los años 2007, 2008, 2009 y 2010.



(a) (b) (c)

(d) (e) (f)

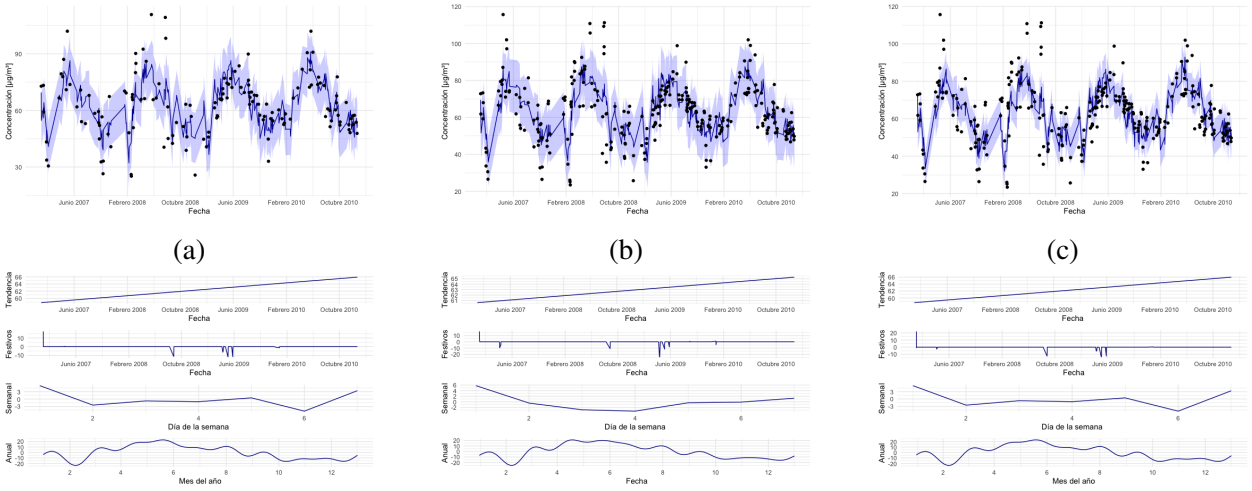
Figura 69: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

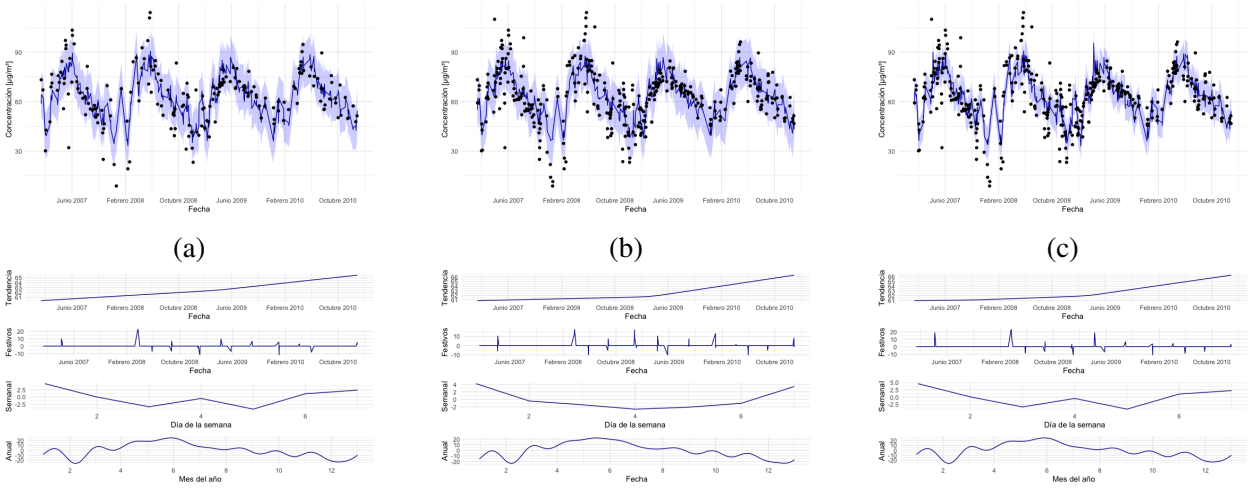
Figura 70: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 71: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

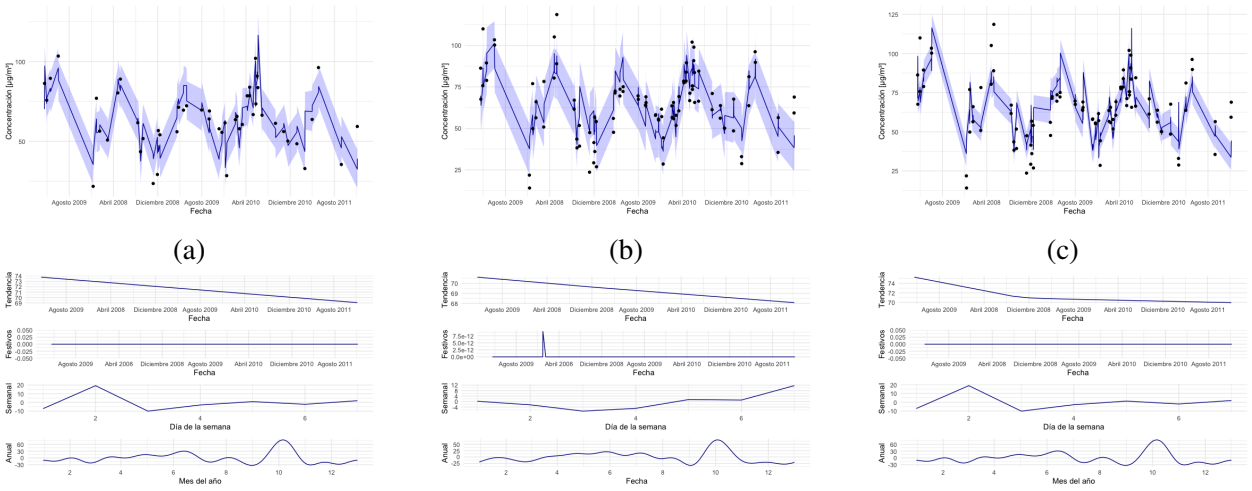


(a) (b) (c)

(d) (e) (f)

Figura 72: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

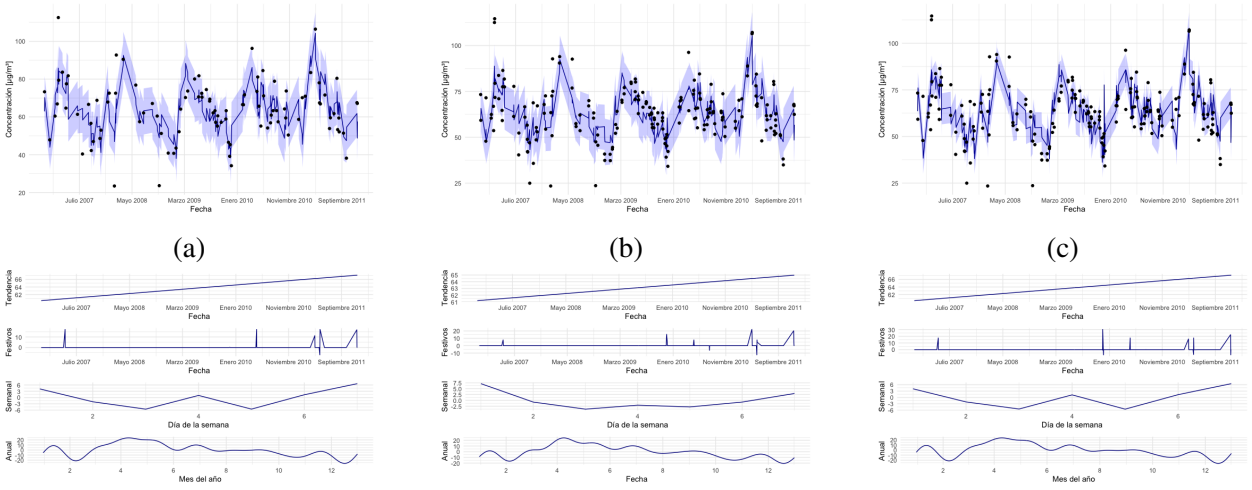
Fechas aleatorizadas de entre las fechas de los años 2007, 2008, 2009, 2010 y 2011.



(a) (b) (c)

(d) (e) (f)

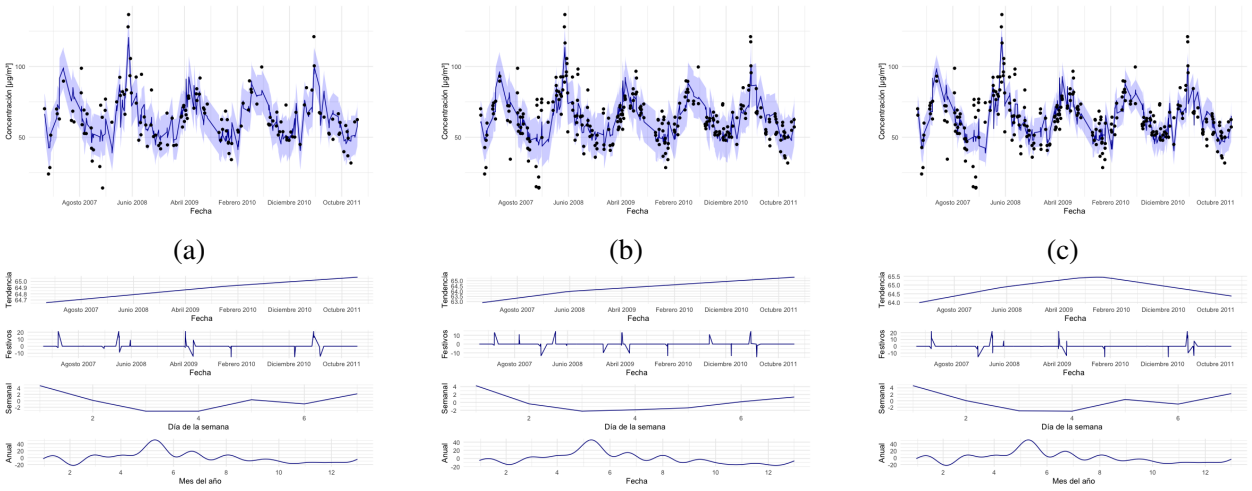
Figura 73: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

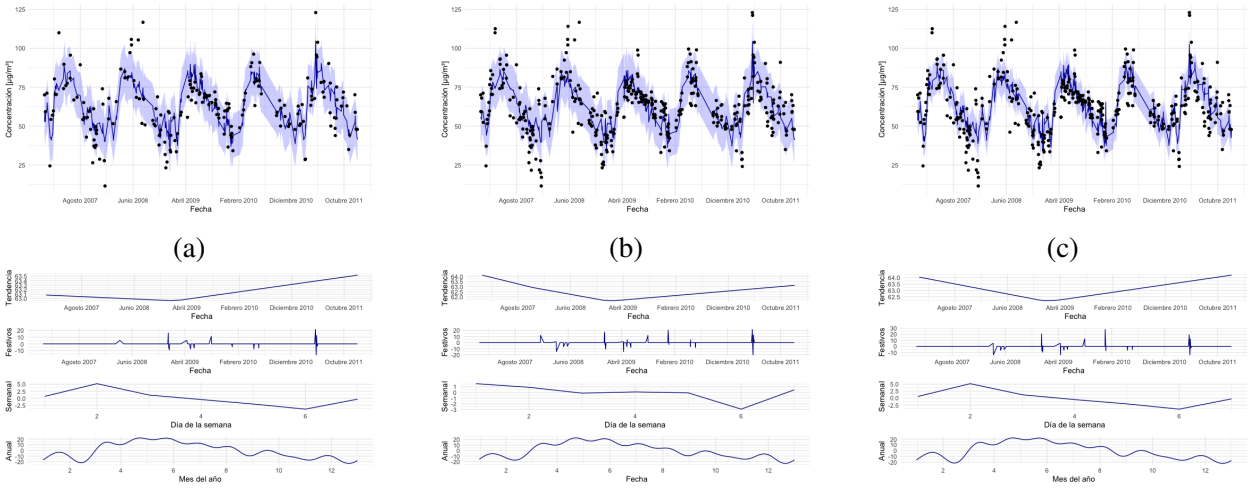
Figura 74: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 75: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

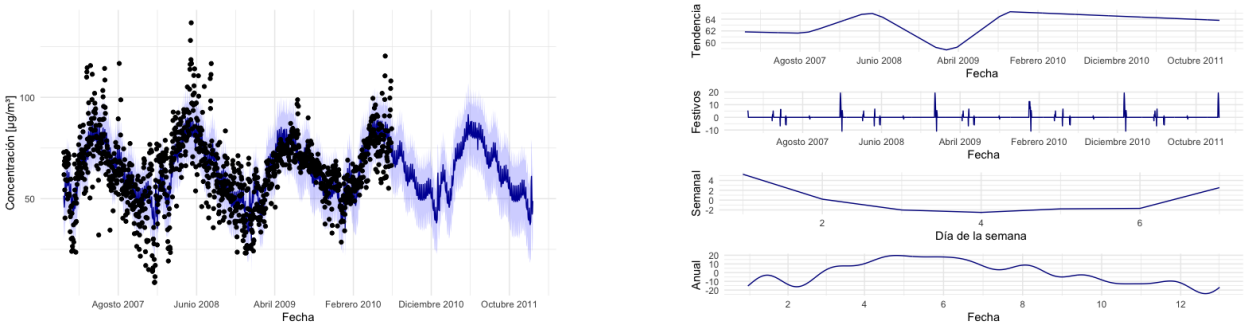


(a) (b) (c)

(d) (e) (f)

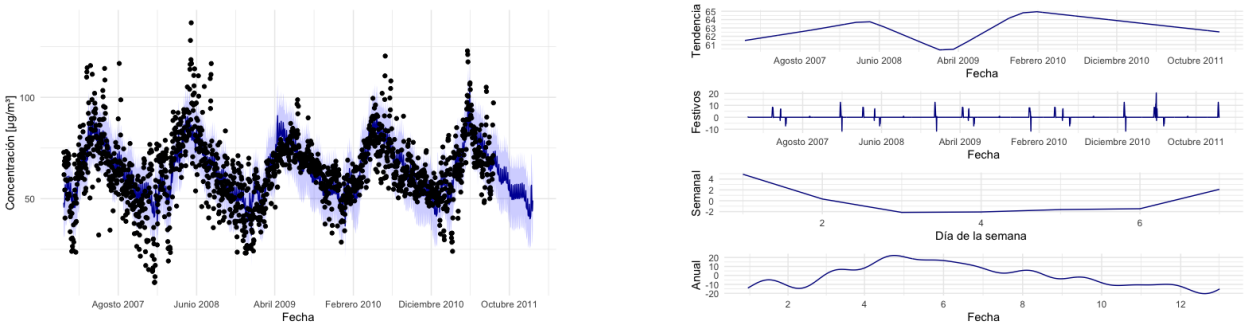
Figura 76: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

Segundo acercamiento



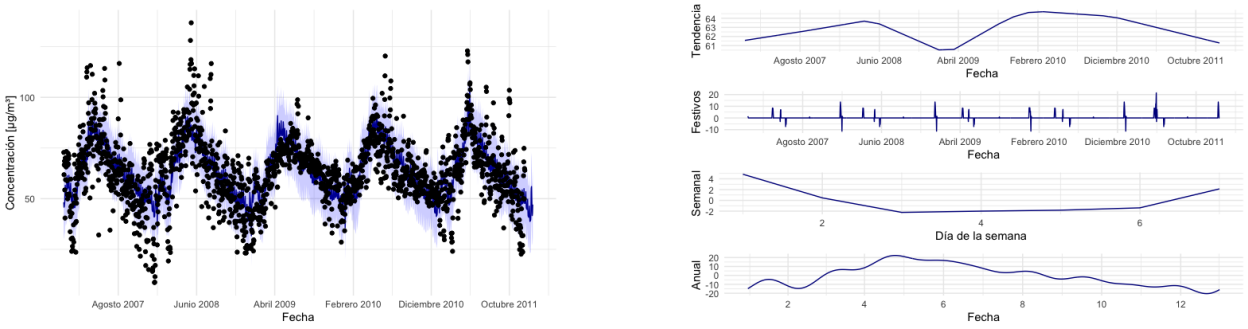
(a) (b)

Figura 77: Predicción del modelo ajustado con los datos de la región 1 del O_3 desde el 2007-01-01 hasta el 2010-07-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.



(a) (b)

Figura 78: Predicción del modelo ajustado con los datos de la región 1 del O_3 desde el 2007-01-01 hasta el 2011-08-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.



(a) (b)

Figura 79: Predicción del modelo ajustado con los datos de la región 1 del O_3 desde el 2007-01-01 hasta el 2011-12-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.

Tabla 12: Errores absolutos medios de las predicciones realizadas utilizando los modelos ajustados. Paso representa el paso en el que se realiza la predicción. El año representa hasta qué año se incluye, partiendo desde 2007, para realizar la selección de días aleatorizados.

		Número de días aleatorizados			
	Paso	50	100	150	200
2007	1	7,212	8,023	9,673	
	2	7,572	8,520	9,454	
	3	8,399	8,968	9,586	
2008	1	7,485	10,271	9,414	9,914
	2	7,872	9,659	9,059	9,760
	3	9,152	10,708	9,496	10,158
2009	1	11,192	8,703	8,337	9,097
	2	7,713	8,740	8,427	8,849
	3	11,435	9,389	9,299	9,226
2010	1	10,548	7,968	9,024	8,696
	2	8,967	7,743	8,686	8,482
	3	12,896	8,585	9,188	8,704
2011	1	9,760	8,153	9,080	9,653
	2	9,989	7,788	8,943	9,337
	3	12,501	8,611	9,287	9,874

Anexo C. Gráficas de las predicciones del algoritmo Prophet para los datos pertenecientes al PM_{10}.

Se presentan las gráficas por orden de año y número de datos utilizados para ajustar los modelos.

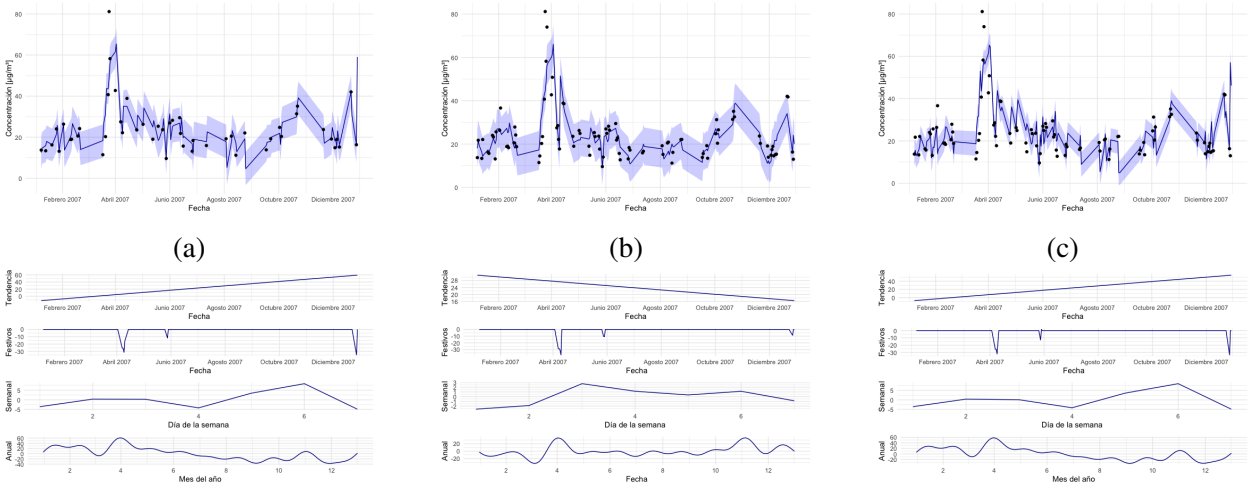
Por cada predicción se presentan dos gráficas:

1. La gráfica de la predicción en la que se muestran los valores predichos en azul oscuro con el rango de error en azul claro y los valores utilizados para ajustar el modelo utilizado para hacer la predicción con puntos negros.

2. La gráfica de las componentes de dicho modelo ajustado, en la que aparecen a su vez cuatro gráficas, en orden de arriba a abajo: Tendencia, efecto de los días festivos, periodicidad semanal, periodicidad anual.

Primer acercamiento

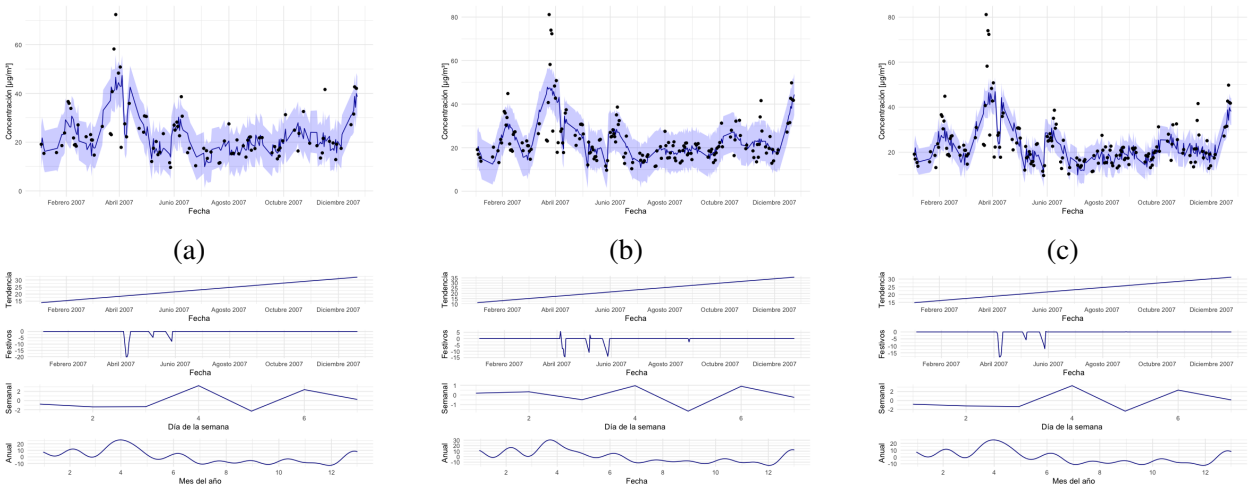
Fechas aleatorizadas de entre las fechas del año 2007.



(a) (b) (c)

(d) (e) (f)

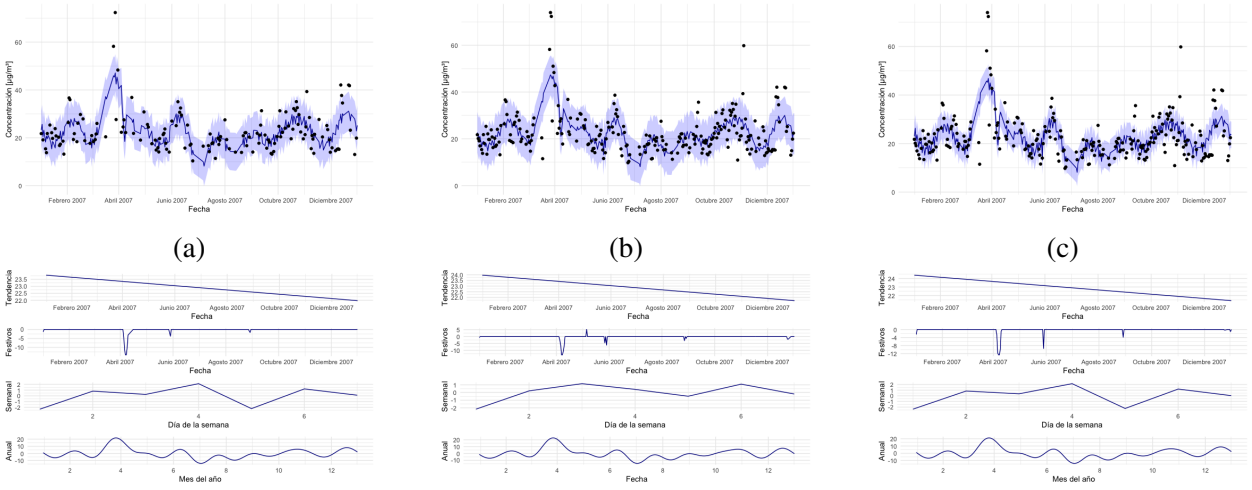
Figura 80: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 81: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

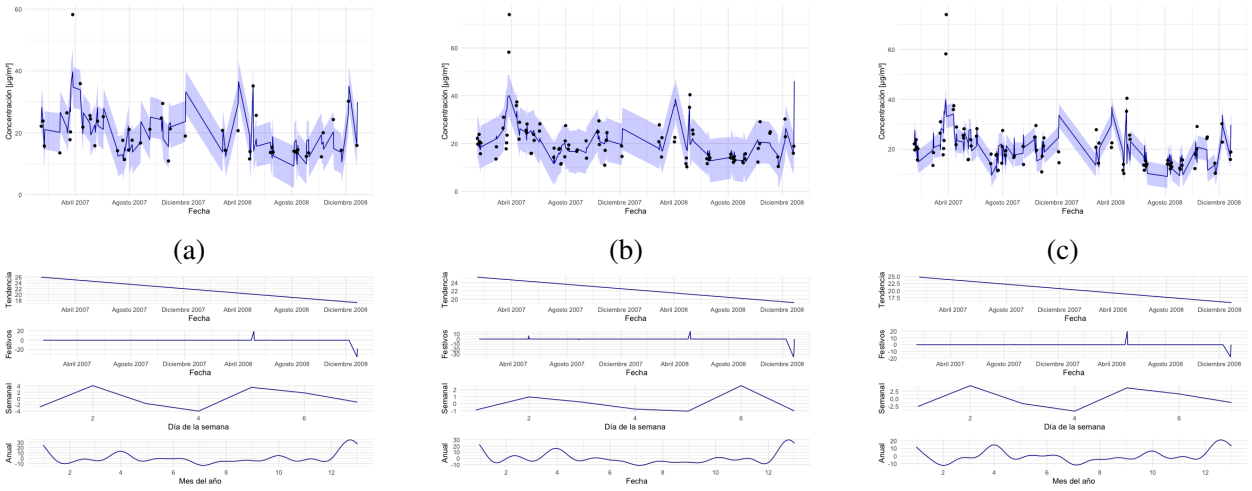


(a) (b) (c)

(d) (e) (f)

Figura 82: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

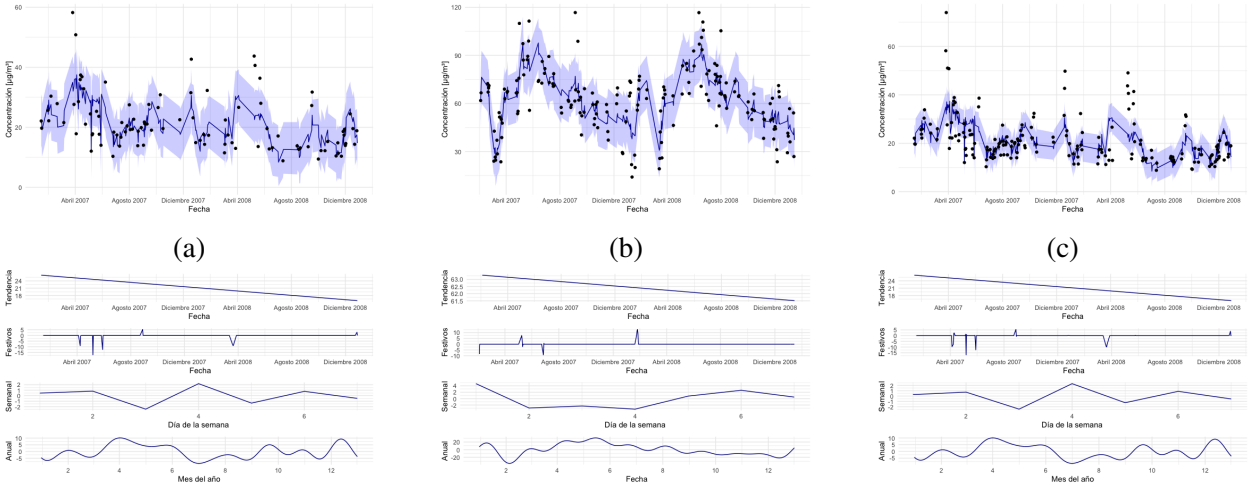
Fechas aleatorizadas de entre las fechas de los años 2007 y 2008.



(a) (b) (c)

(d) (e) (f)

Figura 83: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 84: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

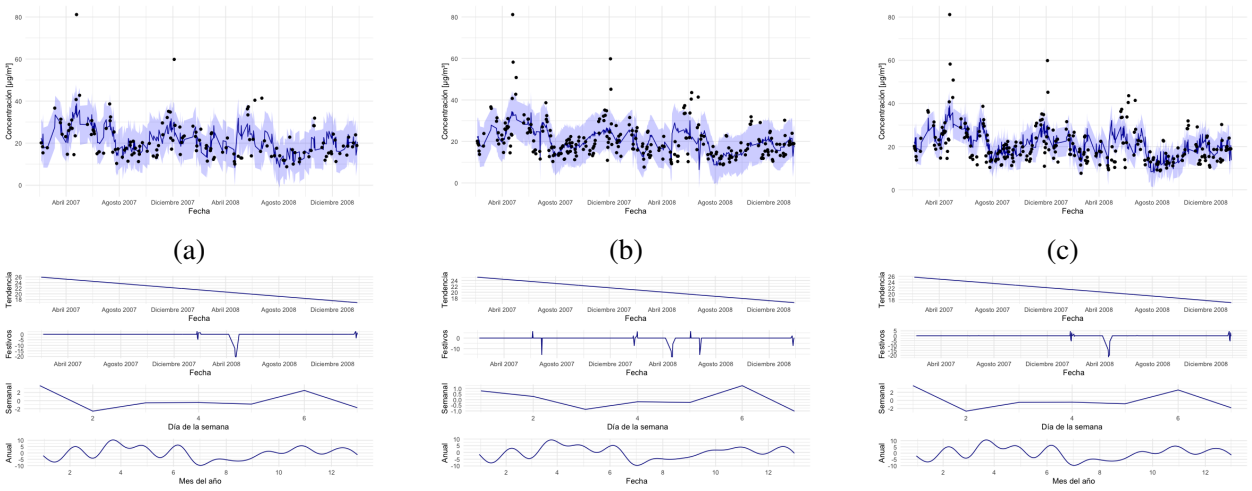


Figura 85: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

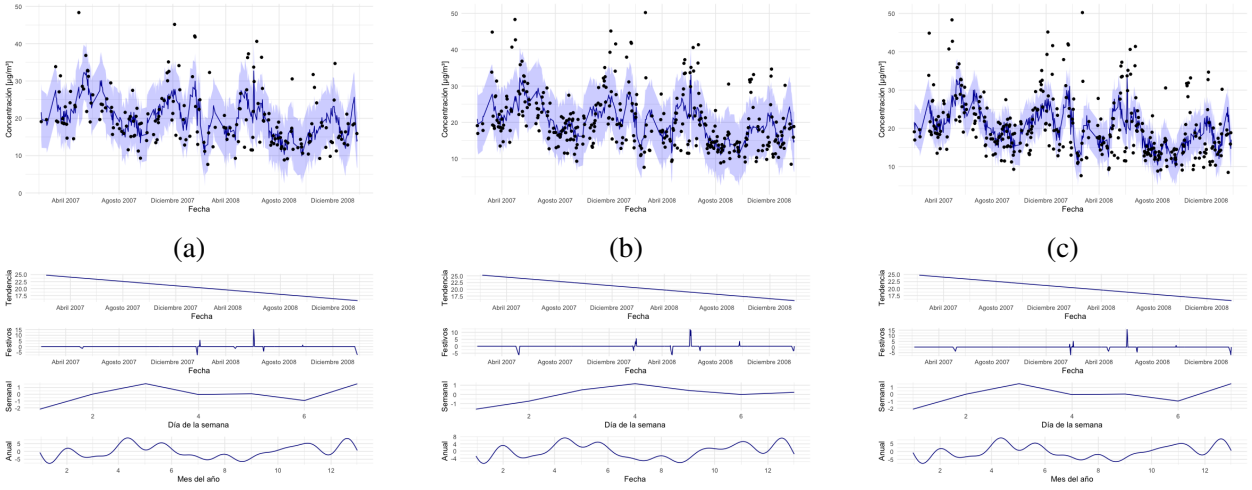


Figura 86: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. De izquierda a derecha paso 1, paso 2 y paso 3. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

Fechas aleatorizadas de entre las fechas de los años 2007, 2008 y 2009.

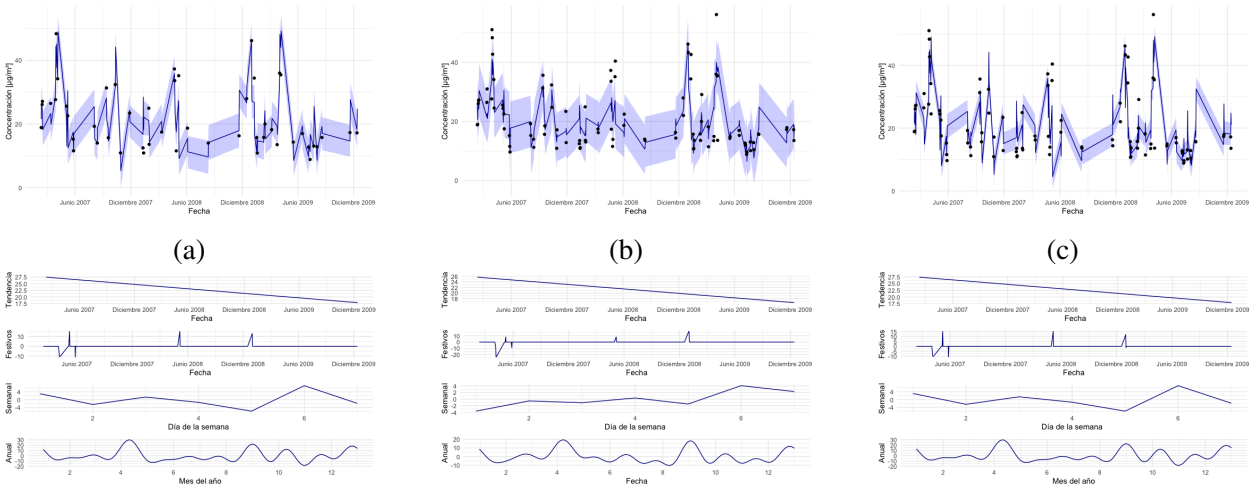


Figura 87: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

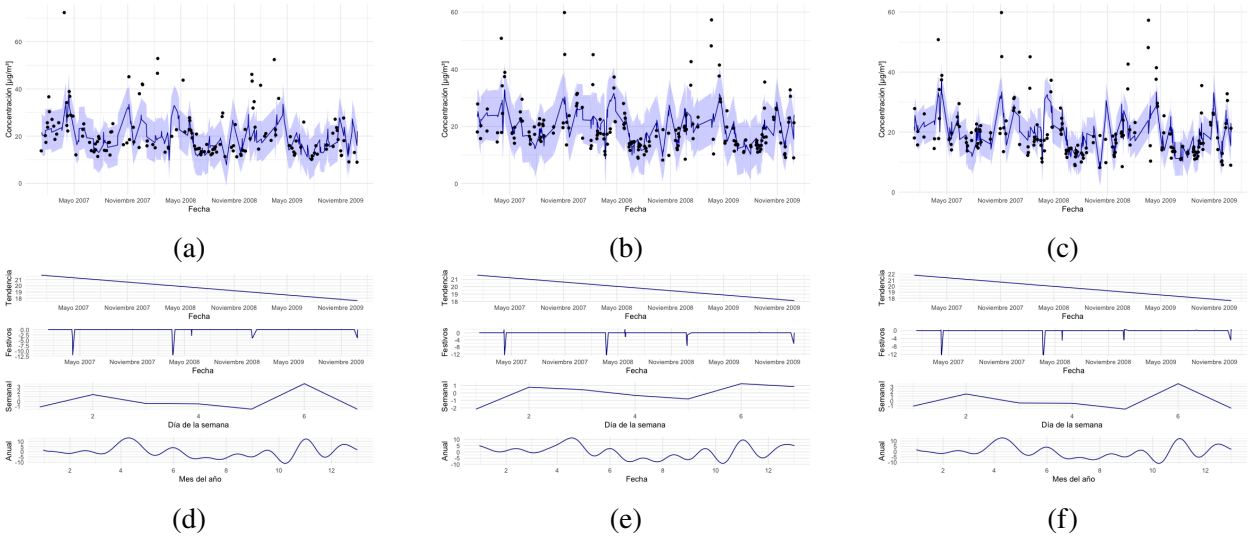
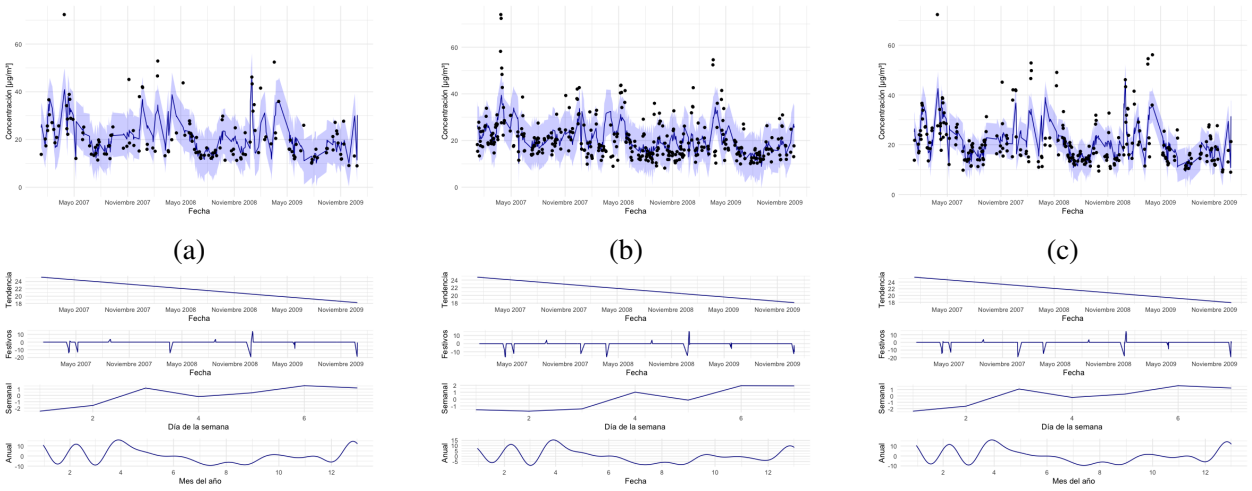


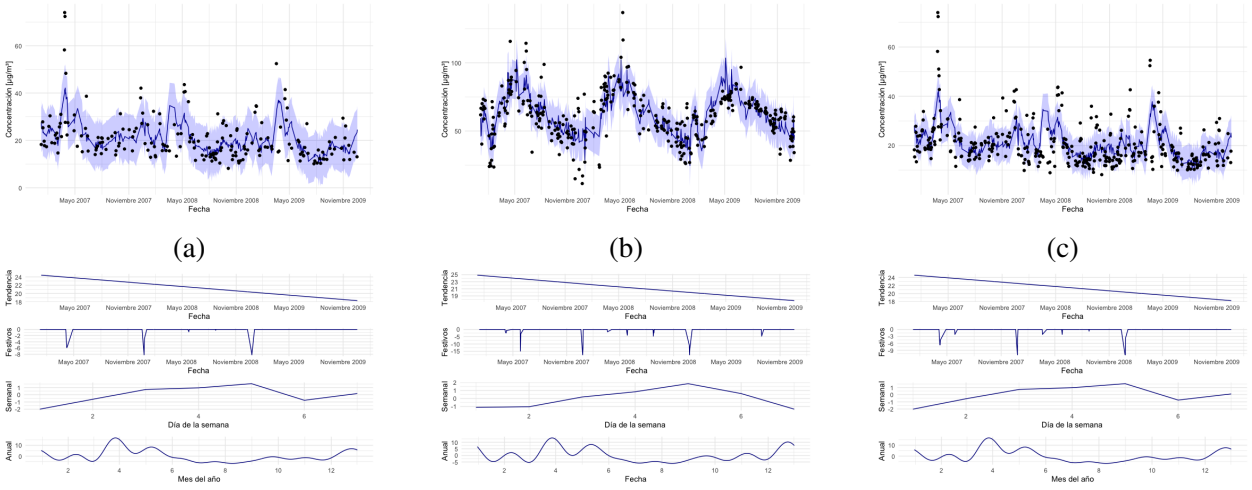
Figura 88: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 89: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

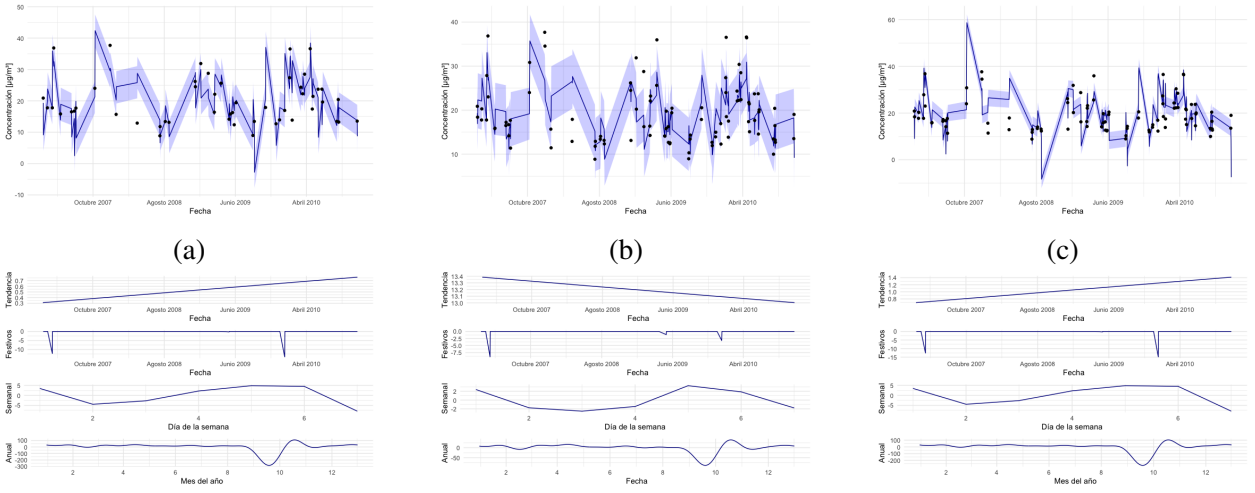


(a) (b) (c)

(d) (e) (f)

Figura 90: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

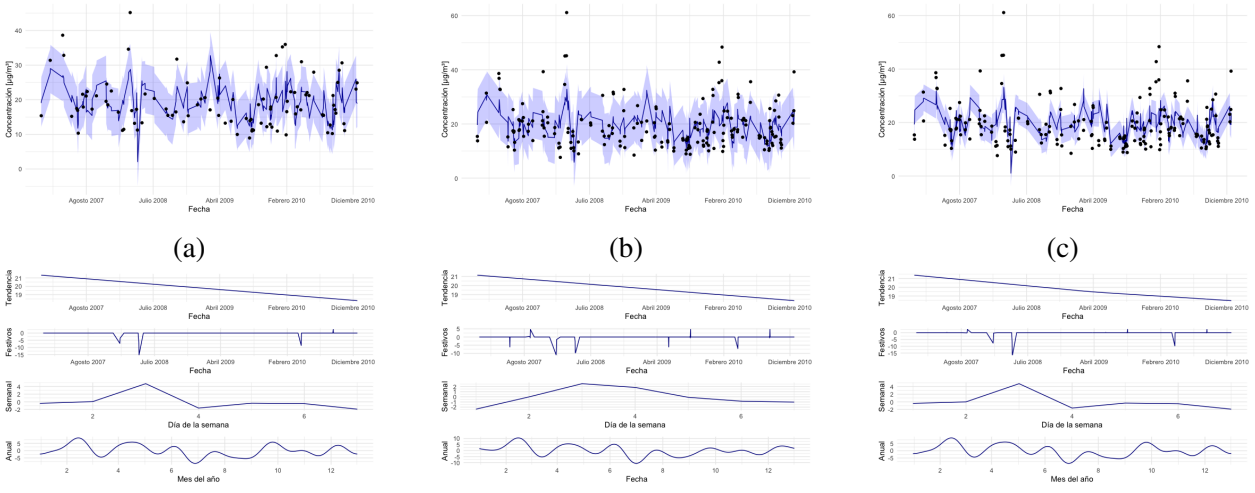
Fechas aleatorizadas de entre las fechas de los años 2007, 2008, 2009 y 2010.



(a) (b) (c)

(d) (e) (f)

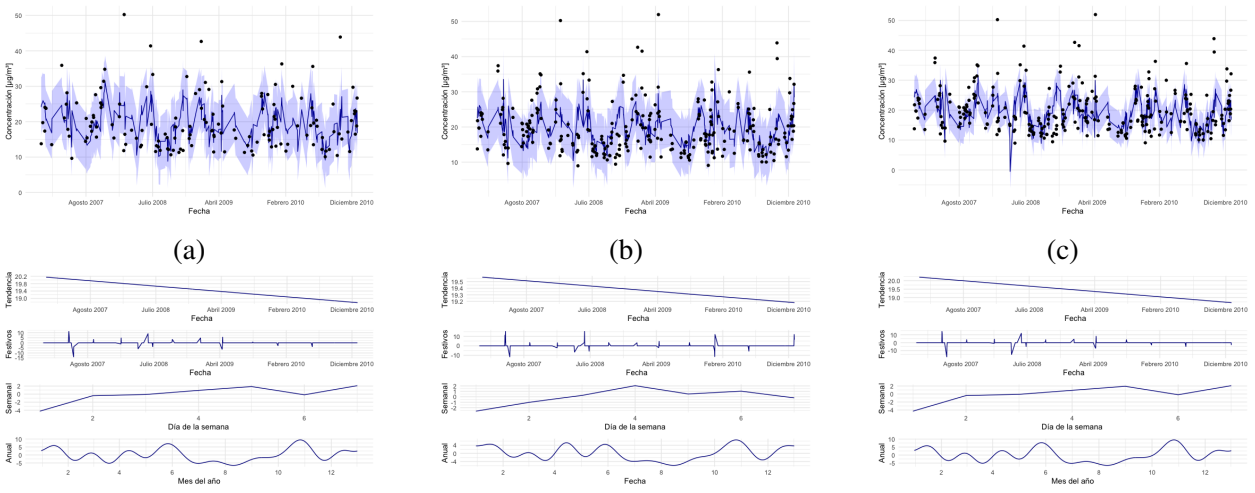
Figura 91: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

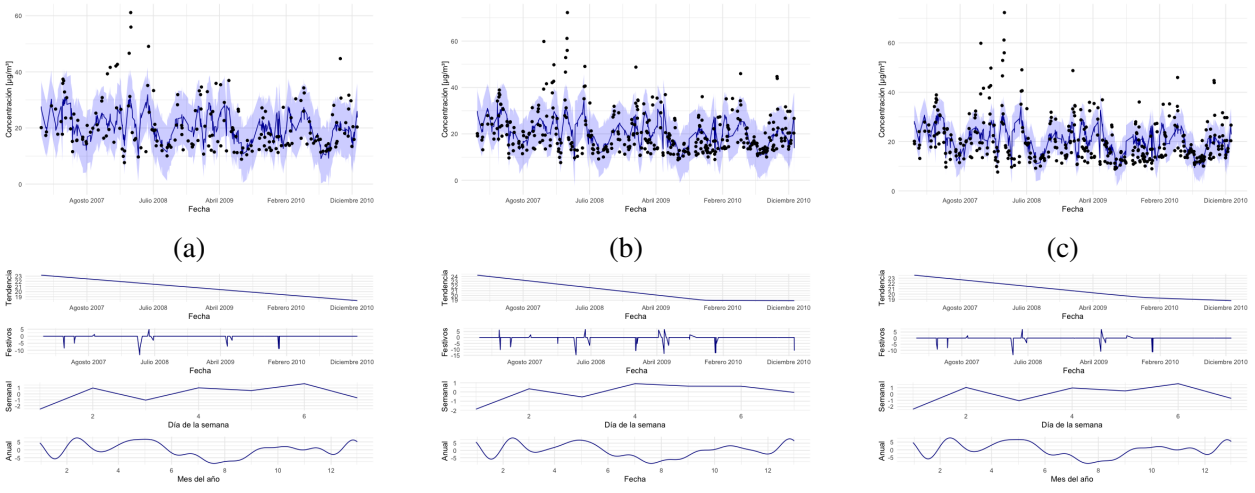
Figura 92: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 93: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

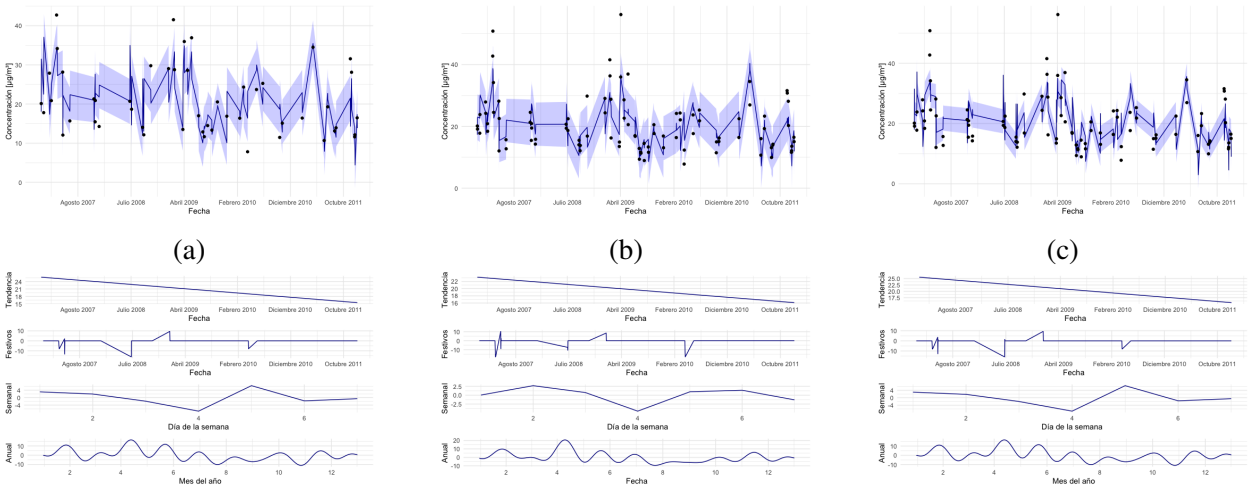


(a) (b) (c)

(d) (e) (f)

Figura 94: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

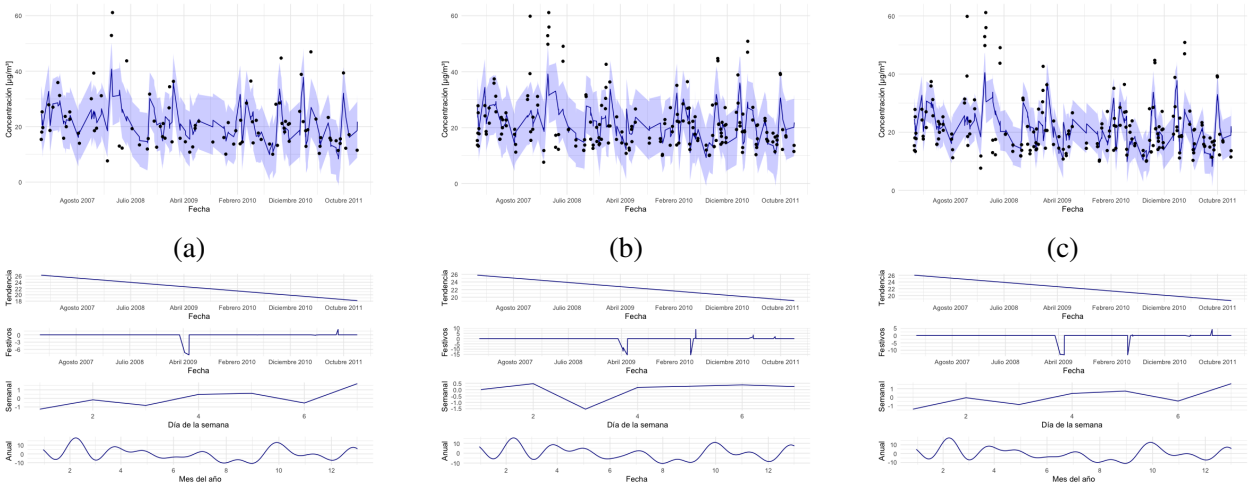
Fechas aleatorizadas de entre las fechas de los años 2007, 2008, 2009, 2010 y 2011.



(a) (b) (c)

(d) (e) (f)

Figura 95: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 96: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

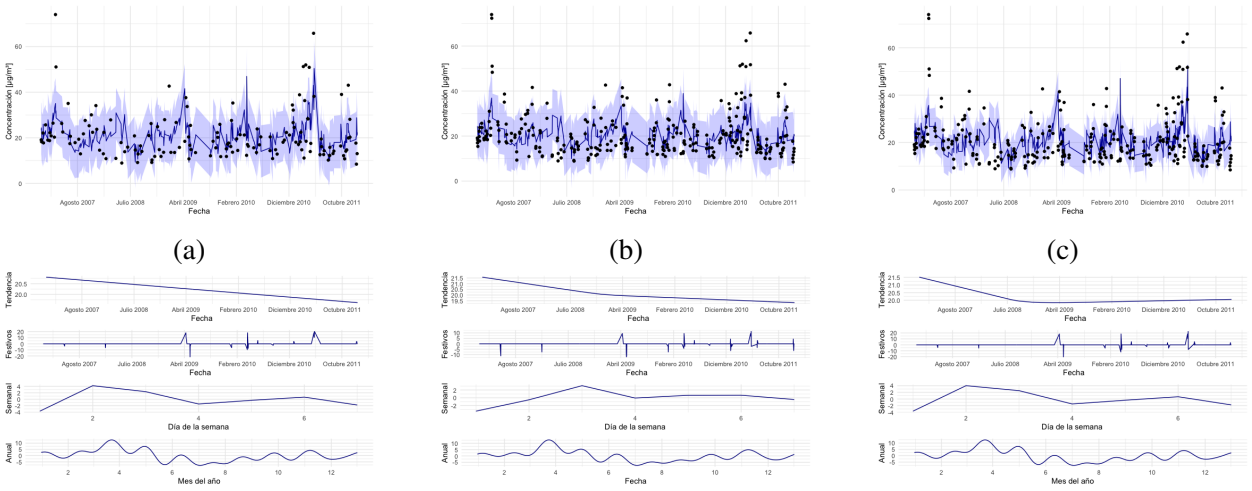


Figura 97: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

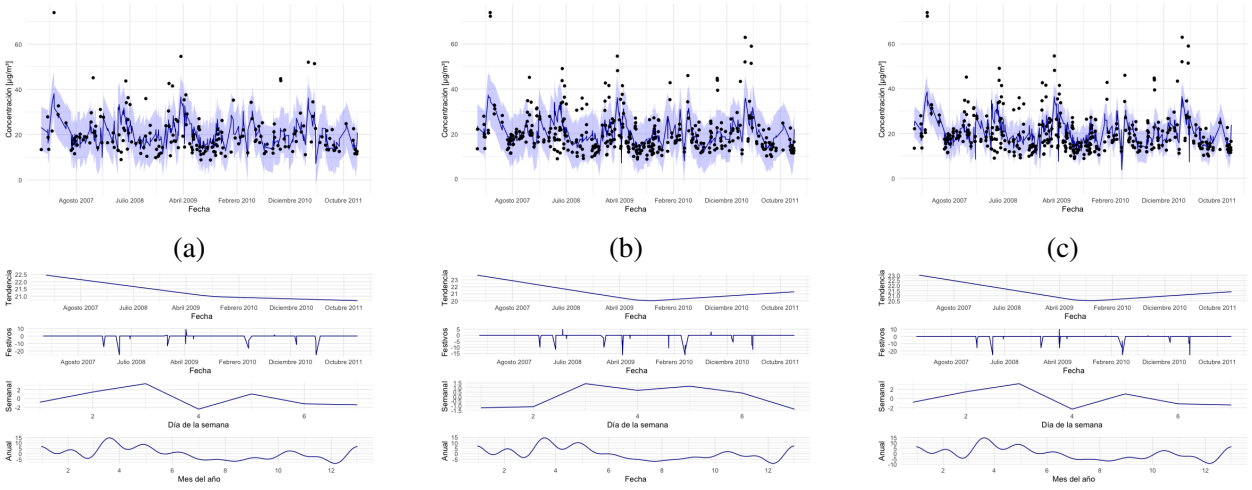
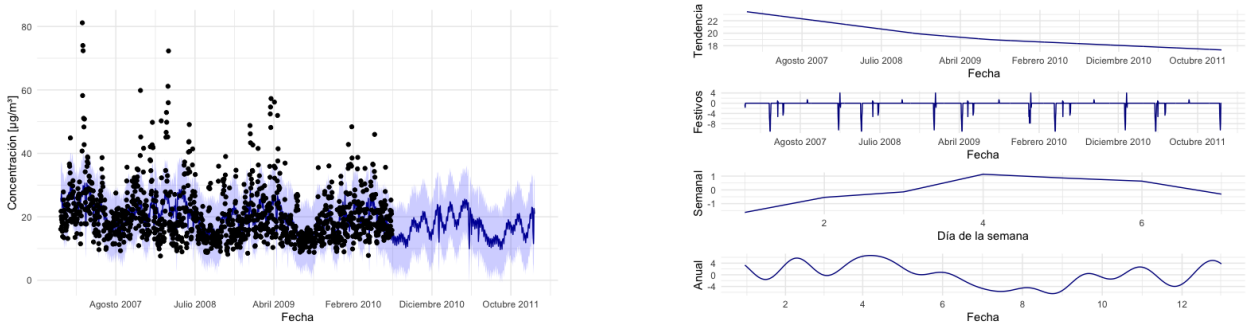


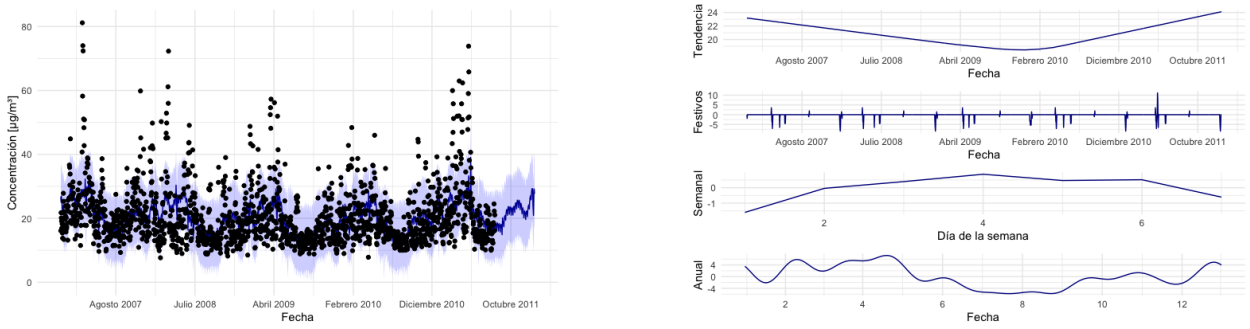
Figura 98: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

Segundo acercamiento



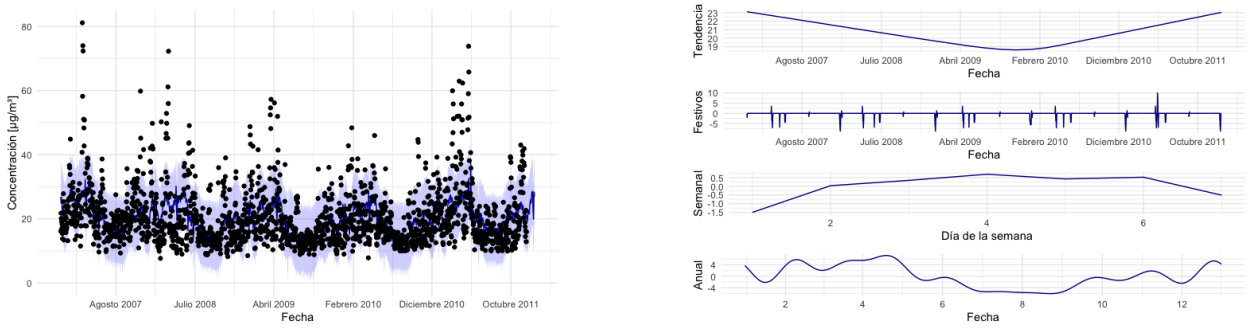
(a) (b)

Figura 99: Predicción del modelo ajustado con los datos de la región 1 de las PM_{10} desde el 2007-01-01 hasta el 2010-07-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.



(a) (b)

Figura 100: Predicción del modelo ajustado con los datos de la región 1 de las PM_{10} desde el 2007-01-01 hasta el 2011-08-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.



(a) (b)

Figura 101: Predicción del modelo ajustado con los datos de la región 1 de las PM_{10} desde el 2007-01-01 hasta el 2011-12-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.

Tabla 13: Errores absolutos medios de las predicciones realizadas utilizando los modelos ajustados. Paso representa el paso en el que se realiza la predicción. El año representa hasta qué año se incluye, partiendo desde 2007, para realizar la selección de días aleatorizados.

		Número de días aleatorizados			
	Paso	50	100	150	200
2007	1	5,901	4,963	5,047	
	2	5,308	4,504	4,836	
	3	7,159	4,717	5,013	
2008	1	5,496	5,131	5,489	4,715
	2	5,141	5,246	5,345	4,945
	3	6,005	5,432	5,778	5,066
2009	1	6,230	4,910	5,266	5,479
	2	5,106	4,839	5,205	5,624
	3	6,694	5,027	5,410	5,875
2010	1	4,947	5,433	5,000	5,835
	2	4,346	5,516	4,837	5,921
	3	6,156	5,716	5,215	6,002
2011	1	5,079	5,983	7,011	6,062
	2	4,828	6,272	6,374	5,832
	3	6,019	6,404	6,783	6,238

Anexo D. Gráficas de las predicciones del algoritmo Prophet para los datos pertenecientes al PM_{25}.

Se presentan las gráficas por orden de año y número de datos utilizados para ajustar los modelos.

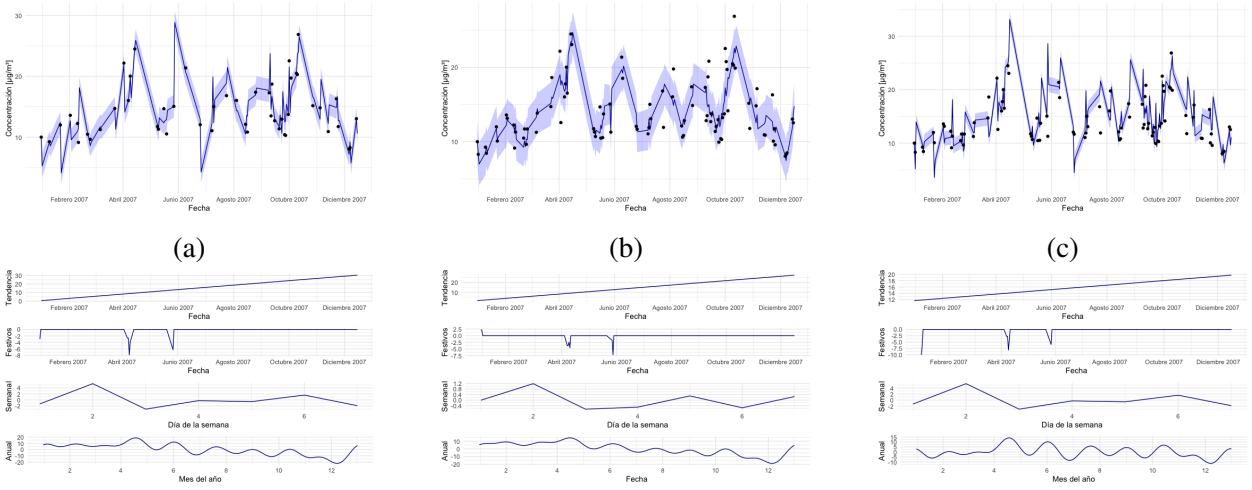
Por cada predicción se presentan dos gráficas:

1. La gráfica de la predicción en la que se muestran los valores predichos en azul oscuro con el rango de error en azul claro y los valores utilizados para ajustar el modelo utilizado para hacer la predicción con puntos negros.

2. La gráfica de las componentes de dicho modelo ajustado, en la que aparecen a su vez cuatro gráficas, en orden de arriba a abajo: Tendencia, efecto de los días festivos, periodicidad semanal, periodicidad anual.

Primer acercamiento

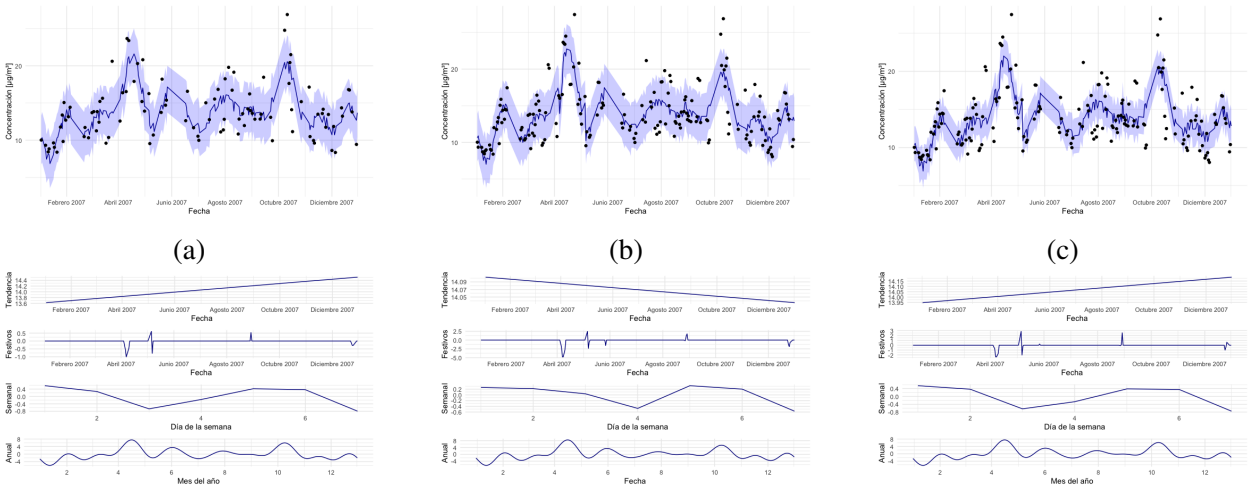
Fechas aleatorizadas de entre las fechas del año 2007.



(a) (b) (c)

(d) (e) (f)

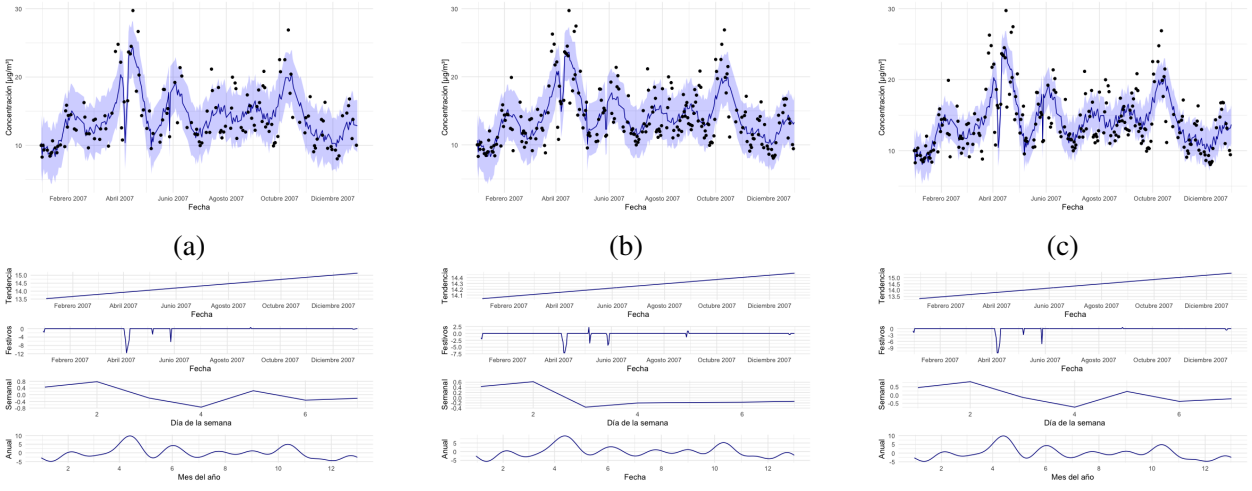
Figura 102: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 103: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

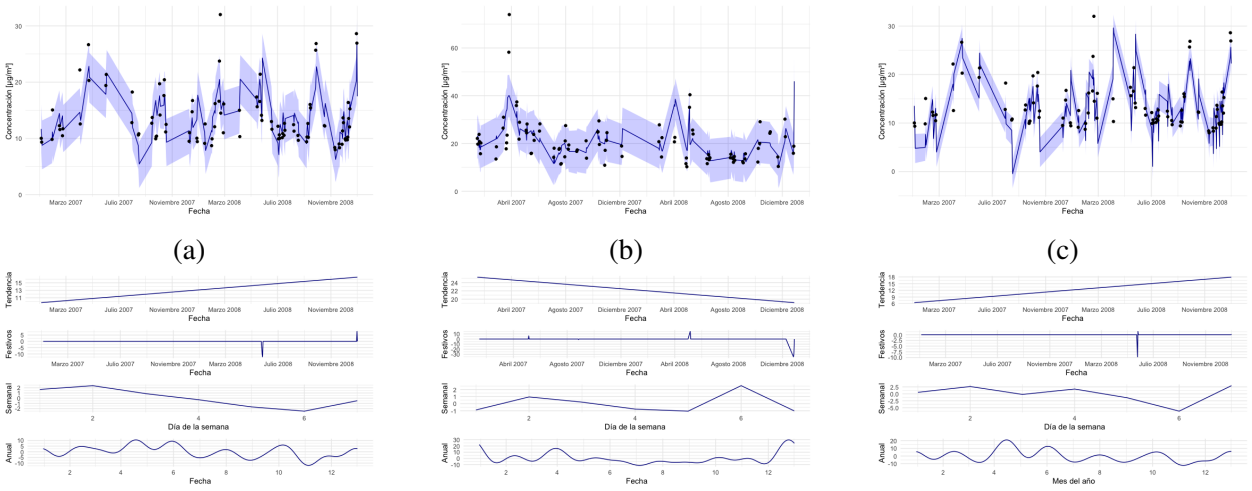


(a) (b) (c)

(d) (e) (f)

Figura 104: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de 2007. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

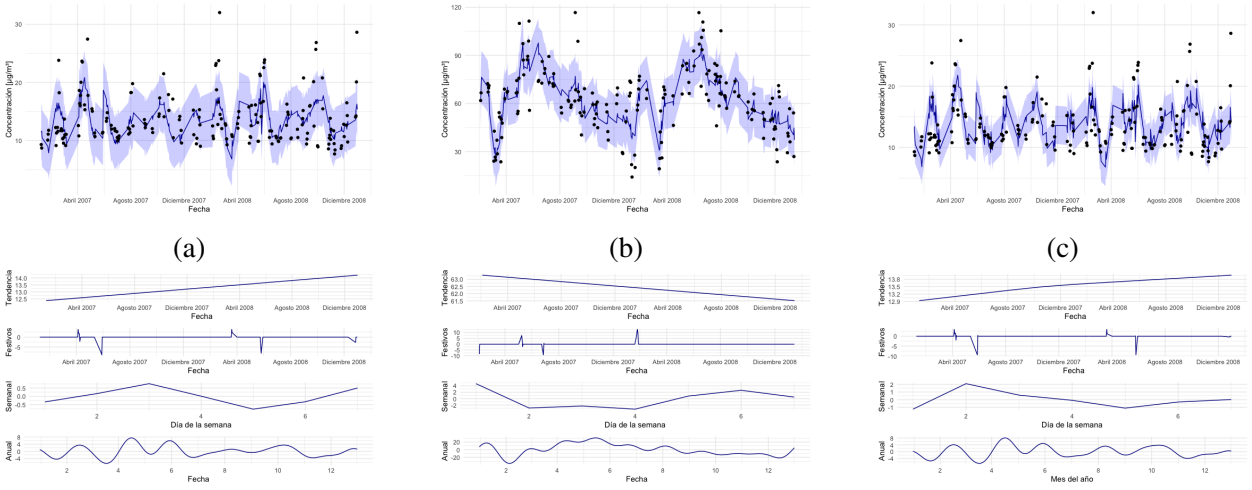
Fechas aleatorizadas de entre las fechas de los años 2007 y 2008.



(a) (b) (c)

(d) (e) (f)

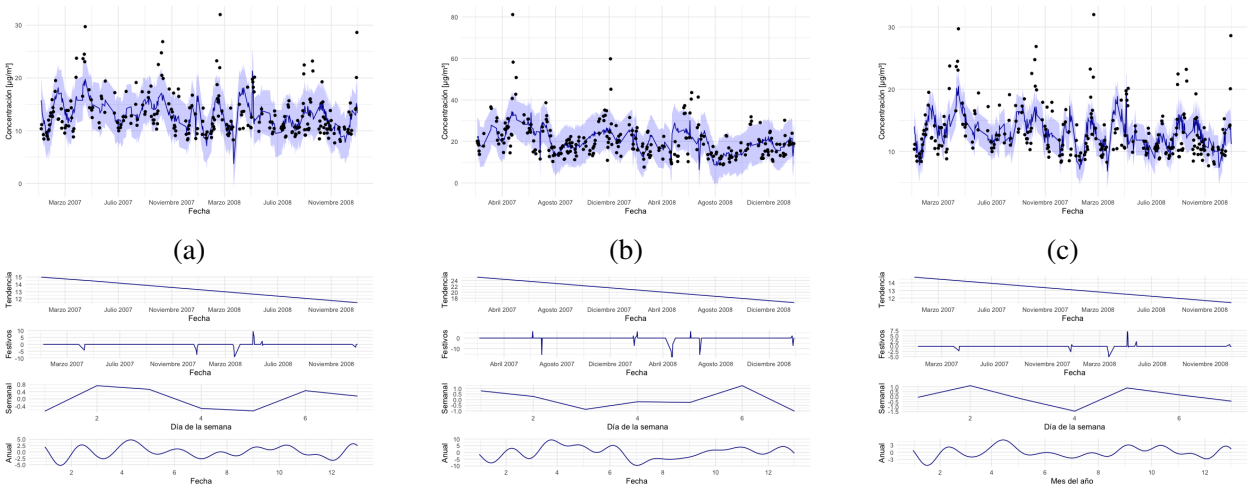
Figura 105: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

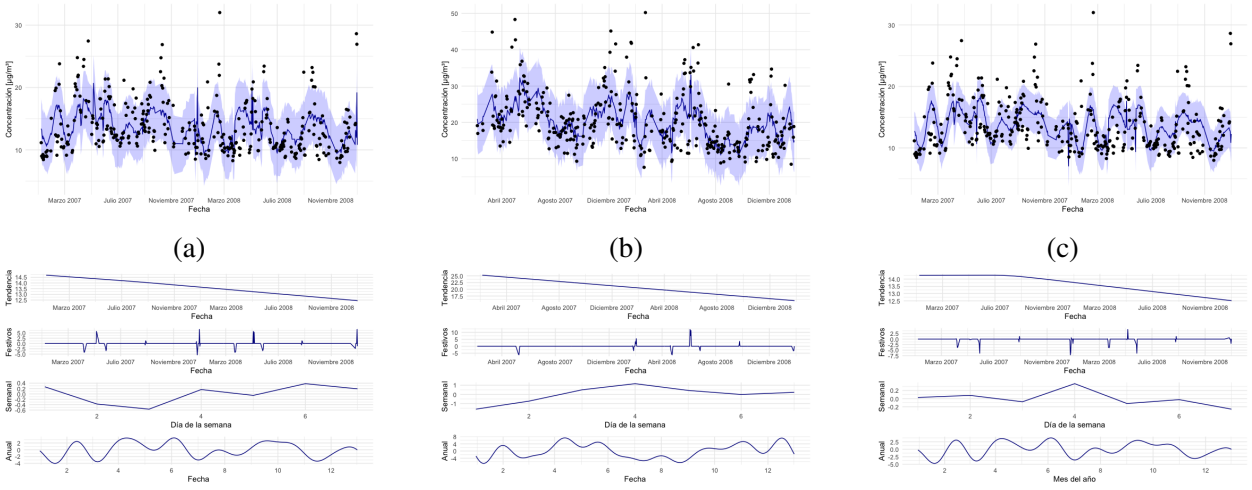
Figura 106: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 107: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

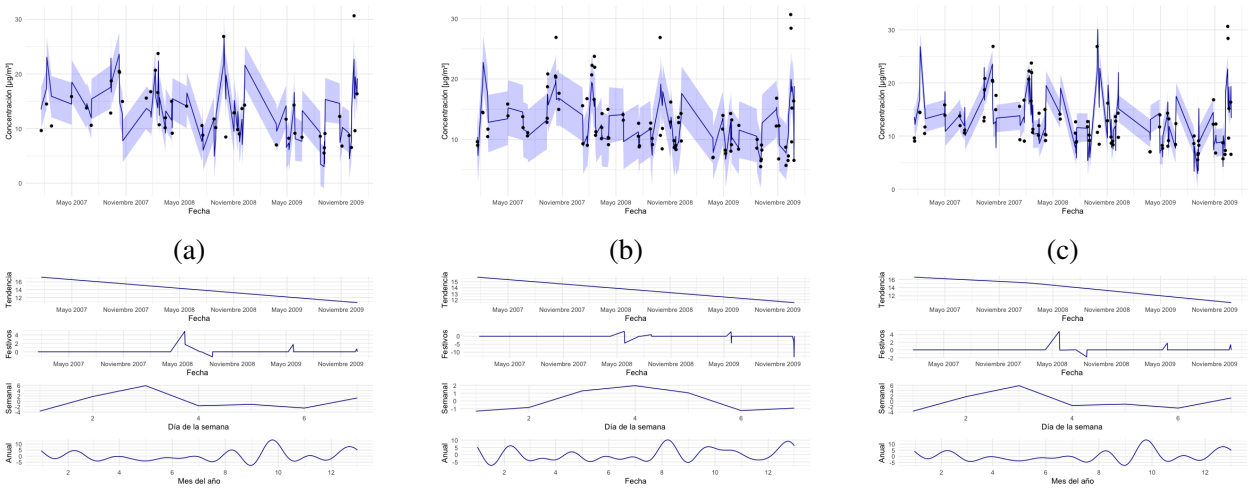


(a) (b) (c)

(d) (e) (f)

Figura 108: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007 y 2008. De izquierda a derecha paso 1, paso 2 y paso 3. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

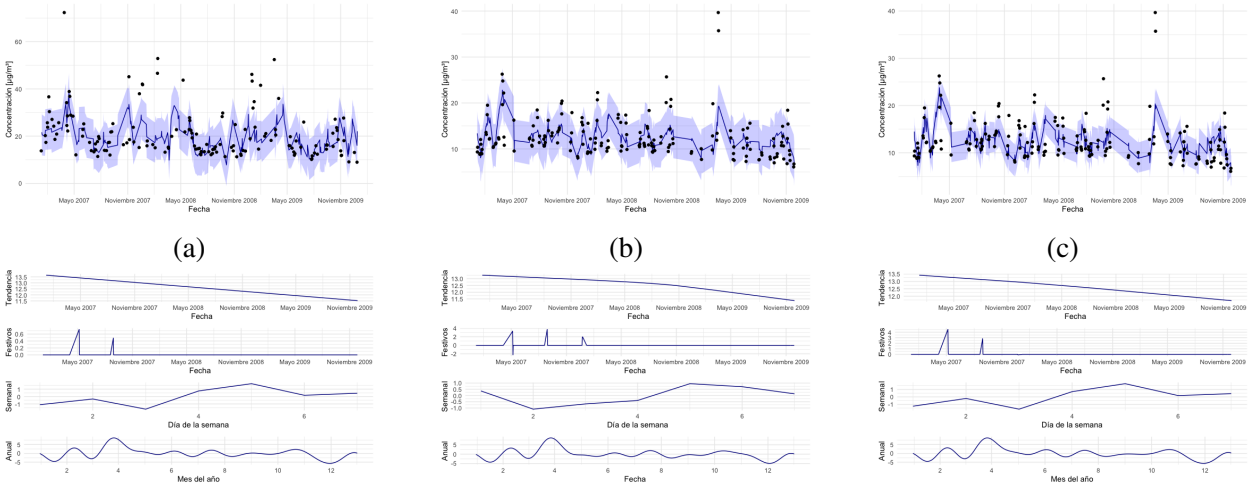
Fechas aleatorizadas de entre las fechas de los años 2007, 2008 y 2009.



(a) (b) (c)

(d) (e) (f)

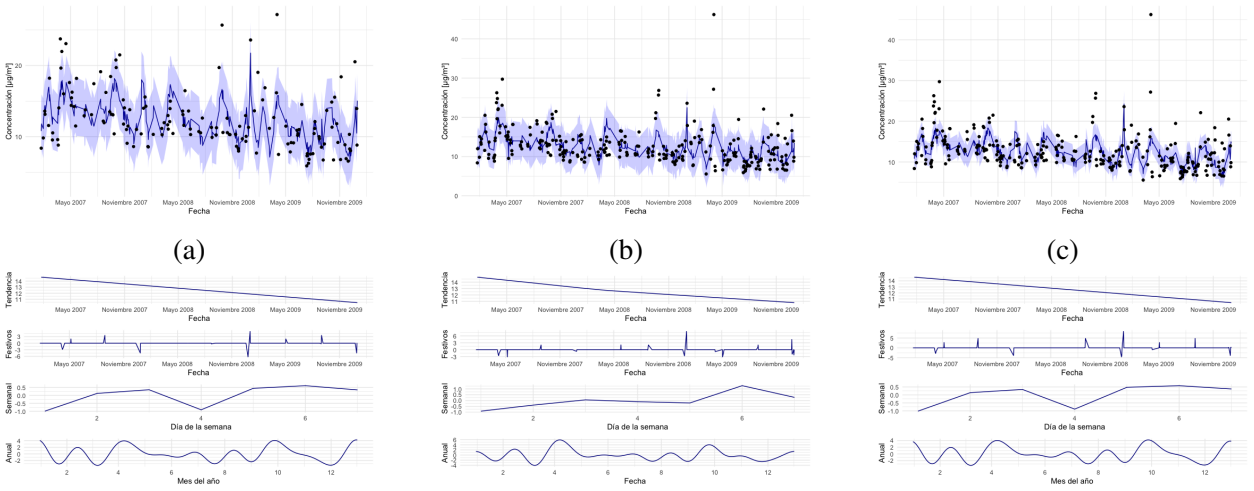
Figura 109: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

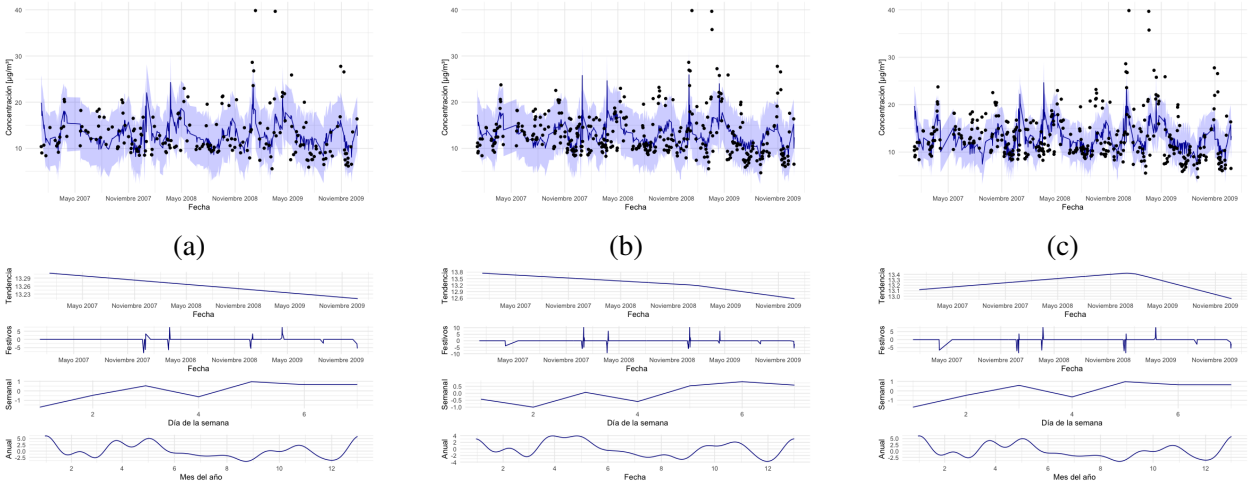
Figura 110: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 111: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

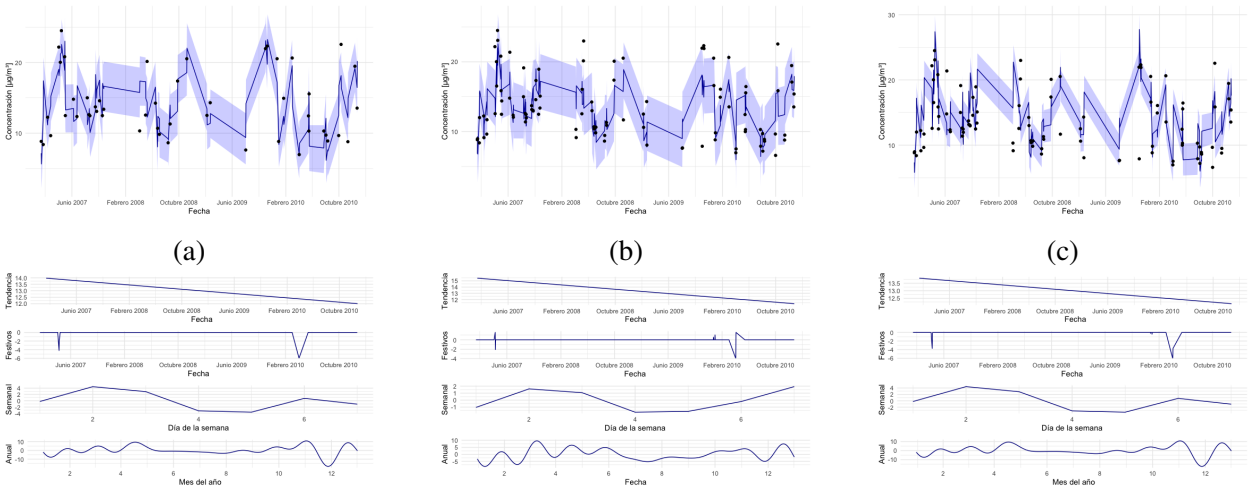


(a) (b) (c)

(d) (e) (f)

Figura 112: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007, 2008 y 2009. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

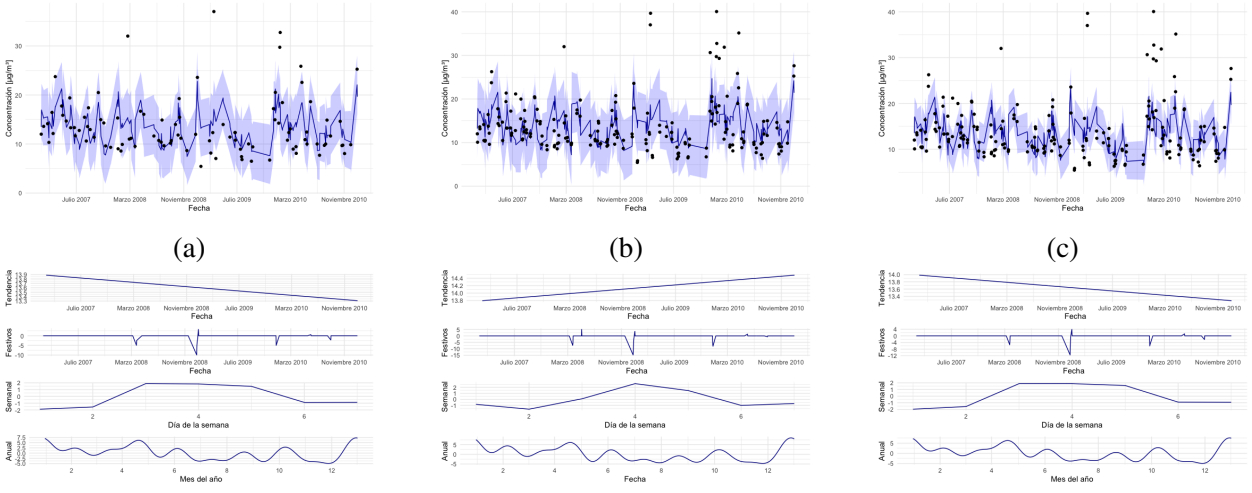
Fechas aleatorizadas de entre las fechas de los años 2007, 2008, 2009 y 2010.



(a) (b) (c)

(d) (e) (f)

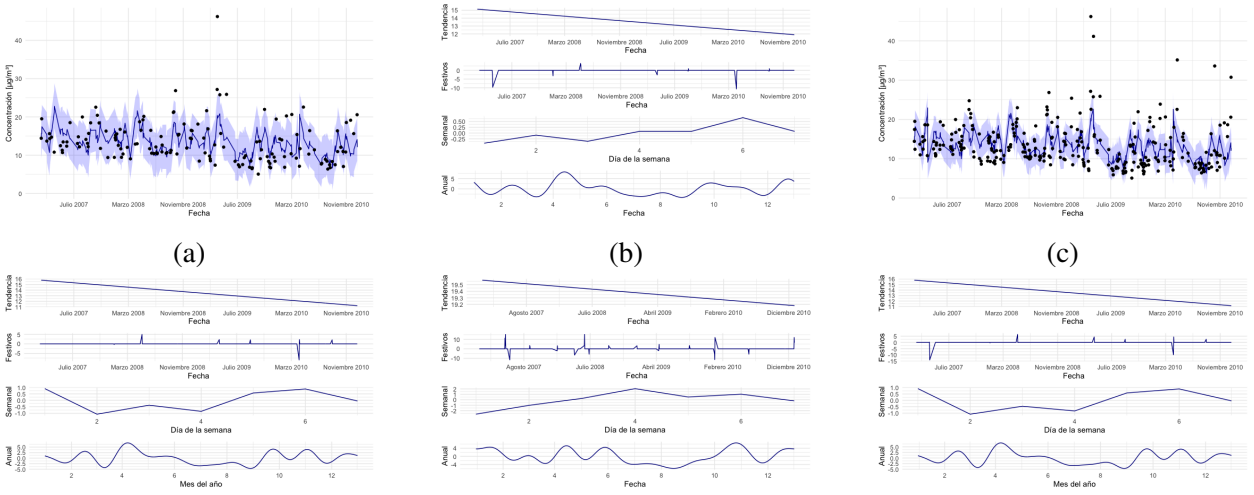
Figura 113: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

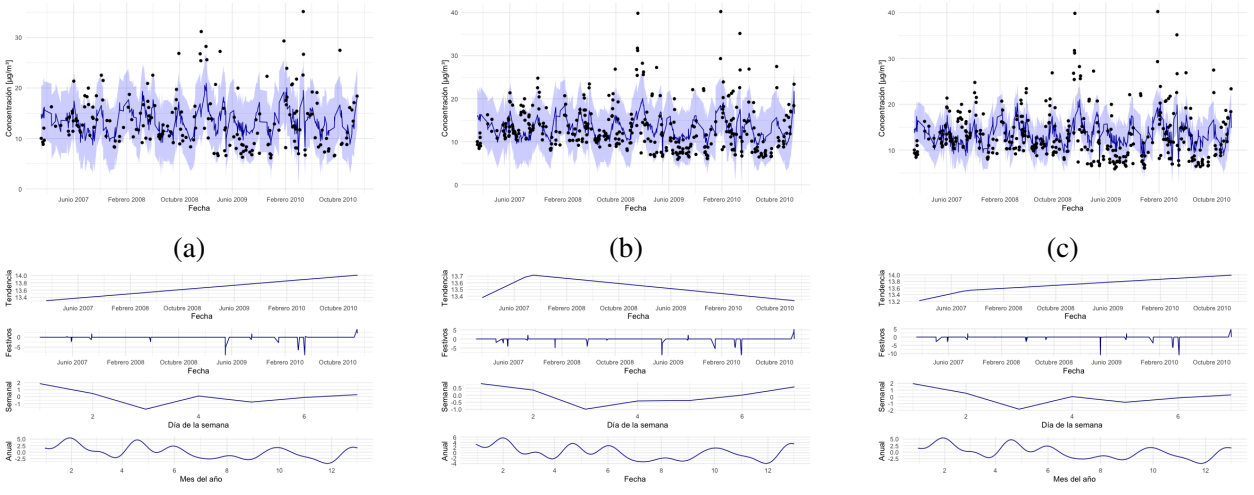
Figura 114: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 115: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

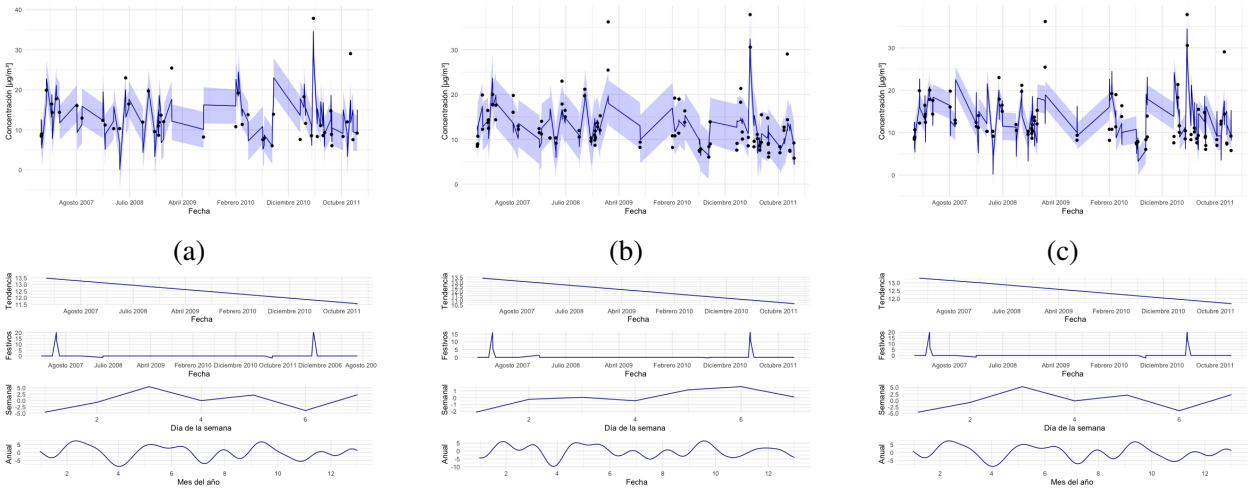


(a) (b) (c)

(d) (e) (f)

Figura 116: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007, 2008, 2009 y 2010. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

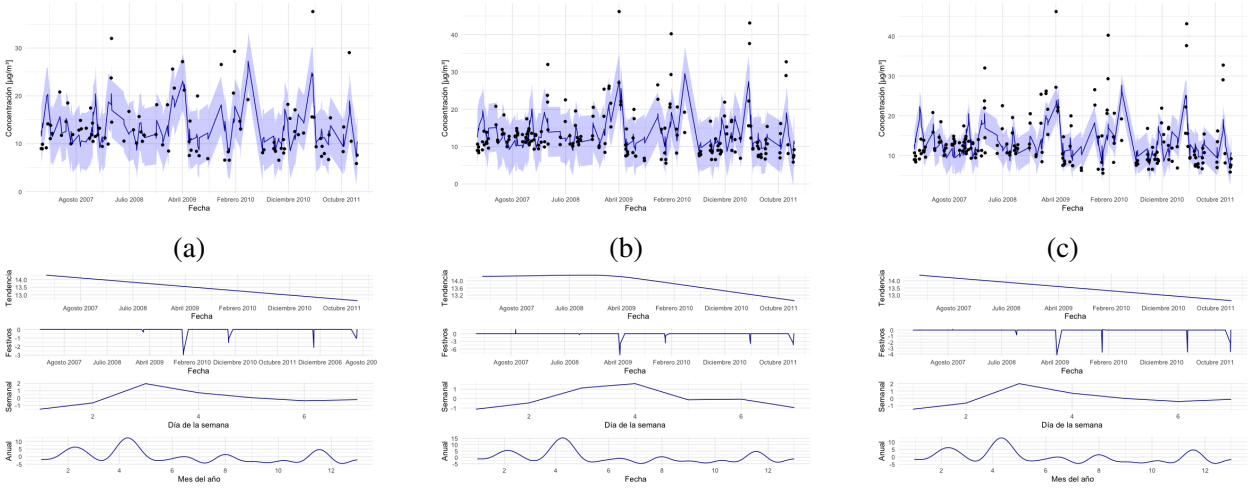
Fechas aleatorizadas de entre las fechas de los años 2007, 2008, 2009, 2010 y 2011.



(a) (b) (c)

(d) (e) (f)

Figura 117: Predicciones realizadas con el modelo ajustado con 50 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.



(a) (b) (c)

(d) (e) (f)

Figura 118: Predicciones realizadas con el modelo ajustado con 100 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

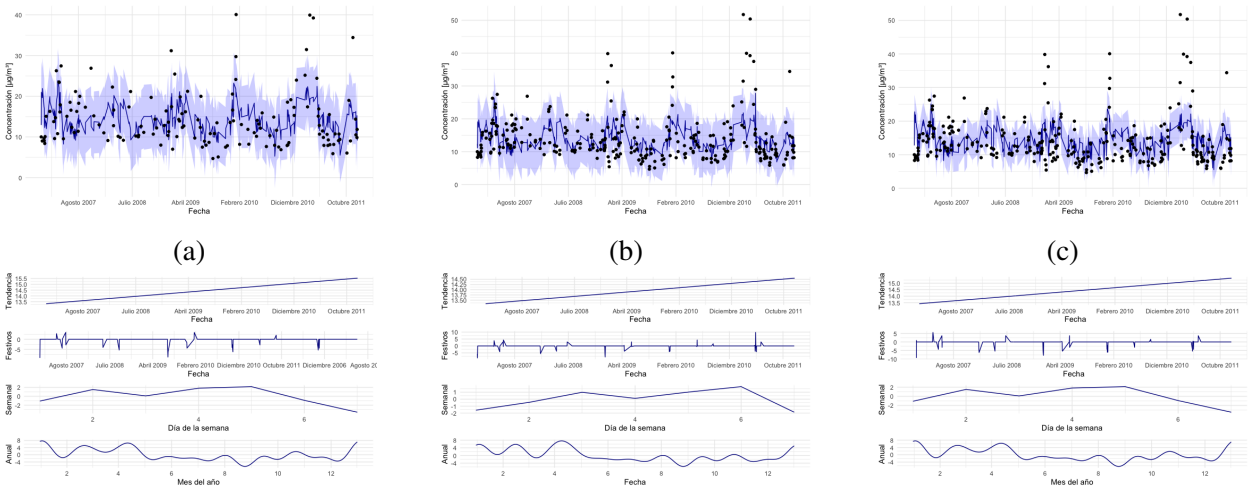


Figura 119: Predicciones realizadas con el modelo ajustado con 150 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

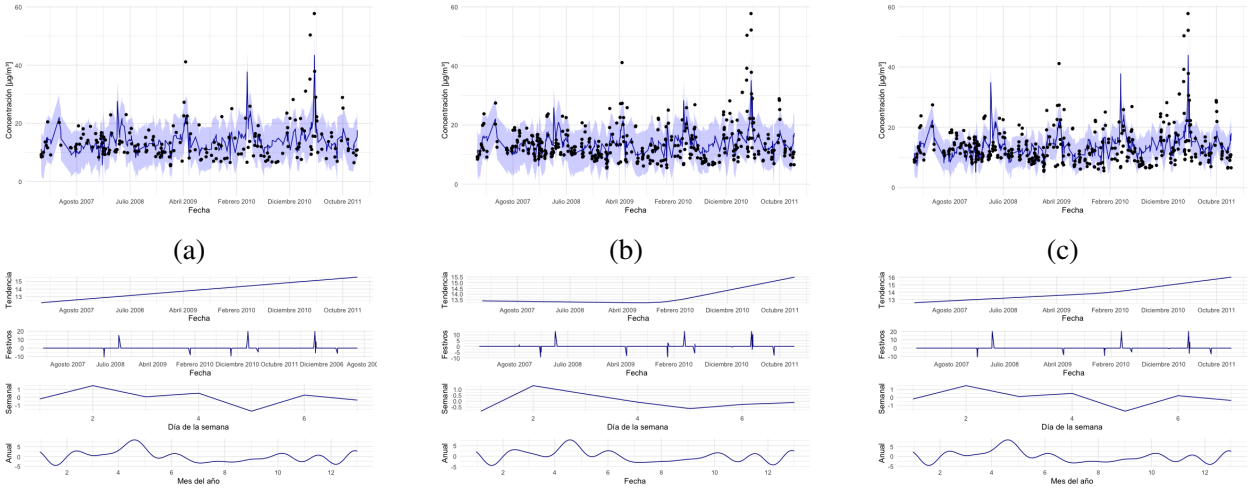
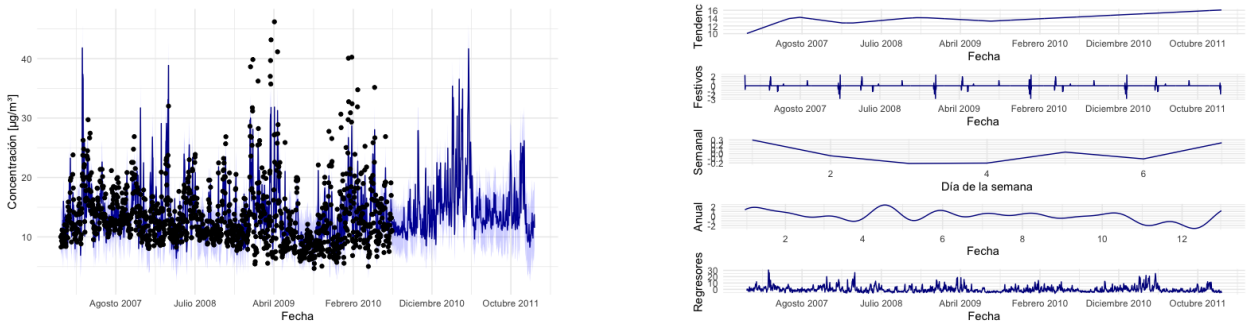


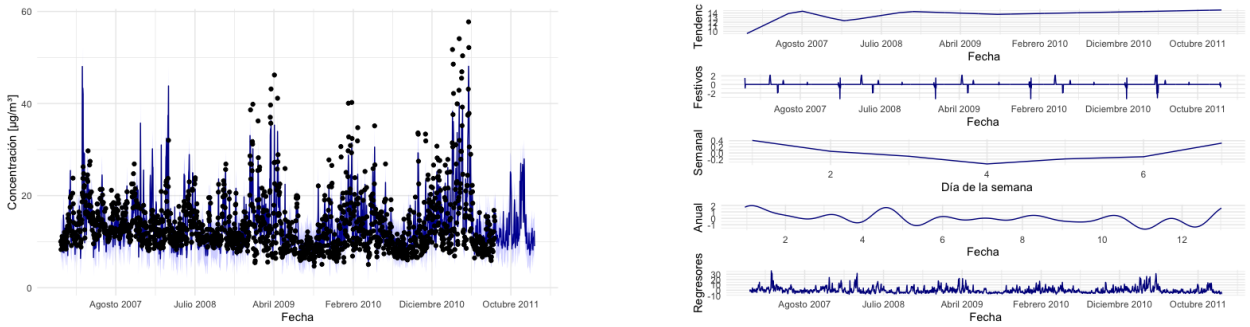
Figura 120: Predicciones realizadas con el modelo ajustado con 200 días elegidos aleatoriamente de entre los días de los años 2007 hasta 2011 incluido. (a) Paso 1. (b) Paso 2. (c) Paso 3. (d) Componentes paso 1. (e) Componentes paso 2. (f) Componentes paso 3.

Segundo acercamiento



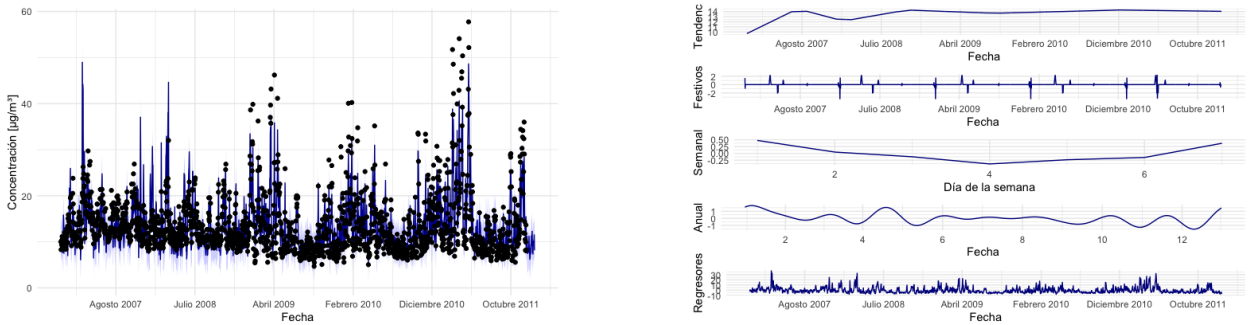
(a) (b)

Figura 121: Predicción del modelo ajustado con los datos de la región 1 de las $PM_{2,5}$ desde el 2007-01-01 hasta el 2010-07-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.



(a) (b)

Figura 122: Predicción del modelo ajustado con los datos de la región 1 de las $PM_{2,5}$ desde el 2007-01-01 hasta el 2011-08-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.



(a) (b)

Figura 123: Predicción del modelo ajustado con los datos de la región 1 de las $PM_{2,5}$ desde el 2007-01-01 hasta el 2011-12-01 sin incluir. Predichos los valores de estas mismas fechas incluyendo las fechas hasta el 2011-12-31 incluido. (a) Predicción. (b) Componentes.

Tabla 14: Errores absolutos medios de las predicciones realizadas utilizando los modelos ajustados. Paso representa el paso en el que se realiza la predicción. El año representa hasta qué año se incluye, partiendo desde 2007, para realizar la selección de días aleatorizados.

		Número de días aleatorizados			
	Paso	50	100	150	200
2007	1	2,533	2,035	2,292	
	2	1,853	2,000	2,152	
	3	3,067	2,034	2,285	
2008	1	3,394	2,928	2,552	2,738
	2	3,060	3,000	2,631	2,721
	3	4,202	3,113	2,724	2,824
2009	1	3,865	2,794	2,712	3,448
	2	2,938	2,803	2,782	3,573
	3	4,209	2,970	2,884	3,744
2010	1	2,998	3,823	3,544	3,737
	2	2,932	4,048	3,588	3,651
	3	3,672	4,016	3,747	3,851
2011	1	4,390	3,930	4,740	4,206
	2	3,324	4,081	4,307	4,620
	3	4,555	4,169	4,771	4,289