DESAFÍO

Rol Data Analyst

Análisis y Predicción de PM10 en Parque O’Higgins (1997-2025)

CANDIDATO:
ELEAZAR ISRAEL MADARIAGA GONZÁLEZ

FECHA
20 / 12 / 2025

CONTENIDOS

1. CONTEXTO

El Sistema de Información Nacional de Calidad del Aire (SINCA) del Ministerio del Medio Ambiente de Chile publica en línea las concentraciones ambientales de los contaminantes atmosféricos actualmente normados, entre ellos el material particulado respirable MP10. Esta información proviene de la red oficial de estaciones de monitoreo, y se utiliza para evaluar el cumplimiento de las normas de calidad del aire y para declarar episodios críticos (alerta, preemergencia y emergencia ambiental).

En este proyecto desarrollado en el marco del desafío propuesto por Datalized para el puesto de Data Analyst, se analizó la serie de concentraciones mensuales de MP10 para la estación Parque O'Higgins, ubicada en la ciudad de Santiago. Esta estación es representativa de la calidad del aire en una zona urbana densamente poblada y con alta actividad vehicular, por lo que sus registros son especialmente relevantes para la evaluación de riesgos en salud y el seguimiento de políticas de descontaminación. Comprender la evolución temporal de las concentraciones de MP10 permite evaluar si las medidas de control han sido efectivas y anticipar periodos en que podrían superarse los niveles de referencia, apoyando así la planificación de medidas preventivas.

1.1. Fuente de información y cobertura temporal

- Fuente: SINCA, https://sinca.mma.gob.cl/index.php/region/index/id/M.
- Formato: archivo descargable .csv, de acceso público.
- Serie utilizada en este trabajo: promedios mensuales de MP10 construidos a partir de promedios móviles diarios.
- Periodo analizado:
    - Registros validados: desde mayo de 1997 hasta julio de 2025.
    - Registros preliminares: agosto, septiembre y octubre de 2025, utilizados sólo para comparar con las predicciones.

1.2. Metodología

En este estudio se construyó una serie temporal utilizando exclusivamente los registros validados. Esta serie se empleó para el análisis descriptivo, la evaluación de estacionariedad (mediante el test de Dickey-Fuller aumentado) y la inspección de la estructura de autocorrelación a través de las funciones ACF y PACF, con el fin de proponer modelos SARIMA candidatos.

Para comparar dichos modelos, la muestra validada se dividió en un conjunto de entrenamiento (desde el inicio de la serie hasta diciembre de 2020) y un conjunto de prueba (enero de 2021 a julio de 2025). Sobre esta partición se estimaron varios

modelos SARIMA y se calcularon, para cada uno, el criterio de información de Akaike (AIC) y el error cuadrático medio (RMSE) en el conjunto de prueba, privilegiando aquellos con buen compromiso entre ajuste y capacidad predictiva fuera de muestra.

Una vez seleccionado el modelo con mejor desempeño según estos criterios, se realizó un diagnóstico de residuos (inspección gráfica, función de autocorrelación y prueba de Ljung–Box) para evaluar la validez de los supuestos del modelo. Posteriormente, el modelo elegido se reestimó utilizando la serie validada completa hasta julio de 2025, y con este modelo final se generaron predicciones para los 12 meses siguientes (de agosto de 2025 a julio de 2026).

Finalmente, las primeras tres predicciones (agosto, septiembre y octubre de 2025) se compararon con las concentraciones preliminares reportadas por SINCA en esos mismos meses. Esta comparación se utilizó como un ejercicio de validación externa a muy corto plazo, calculando el error de predicción e interpretando los resultados a la luz del carácter preliminar de dichos registros.

2. OBJETIVOS

El objetivo principal de este trabajo es modelar la serie mensual de concentraciones de MP10 en la estación Parque O'Higgins y evaluar un modelo SARIMA que permita:

- Describir la dinámica temporal de la serie (tendencias y estacionalidad).
- Evaluar la capacidad predictiva del modelo mediante un conjunto de prueba.
- Generar pronósticos de corto plazo (12 meses) posteriores al último dato validado disponible.
- Validar las predicciones con los registros preliminares más recientes reportados por SINCA.

2.1. Variable objetivo

La variable objetivo corresponde a:

- Nombre: concentración mensual de MP10.
- Unidad: μg/m³N (microgramos por metro cúbico normal).

3. RESULTADOS

3.1. Análisis exploratorio

Antes de especificar un modelo SARIMA, se realizó un análisis exploratorio de la serie mensual de MP10 en la estación Parque O'Higgins utilizando únicamente los datos validados hasta julio de 2025. Este paso es fundamental porque permite entender la estructura básica de la serie (tendencias, estacionalidad, variabilidad) y orientar la elección de los órdenes del modelo.

3.1.1. Visualización de la serie temporal

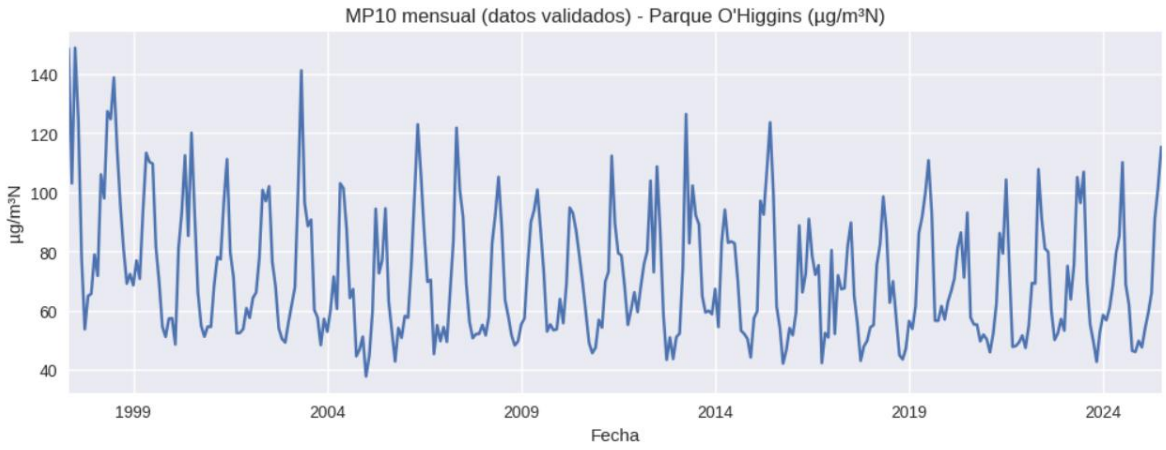

Figura 1. Visualización exploratoria serie de tiempo de estudio.

En primer lugar, se graficó la serie de concentraciones mensuales de MP10 (µg/m³N). La serie muestra una alta variabilidad entre meses, con episodios de concentraciones elevadas y periodos de niveles más moderados, lo que es consistente con la dinámica de la contaminación por material particulado en un entorno urbano. Visualmente se aprecia una estructura estacional: ciertos meses del año tienden a presentar valores sistemáticamente más altos (asociados a condiciones meteorológicas desfavorables para la dispersión de contaminantes), mientras que otros meses muestran niveles relativamente más bajos.

Aunque a lo largo del periodo pueden existir cambios en los niveles promedio asociados a políticas de descontaminación o variación en las emisiones, la serie no presenta una tendencia determinista clara y sostenida que justifique, por sí sola, la incorporación de una tendencia explícita en el modelo. Esta primera visualización sugiere que un modelo de tipo SARIMA (con componente estacional) es una alternativa razonable para capturar la dinámica observada.

3.1.2. Estructura de Auto-Correlación (ACF y PACF)

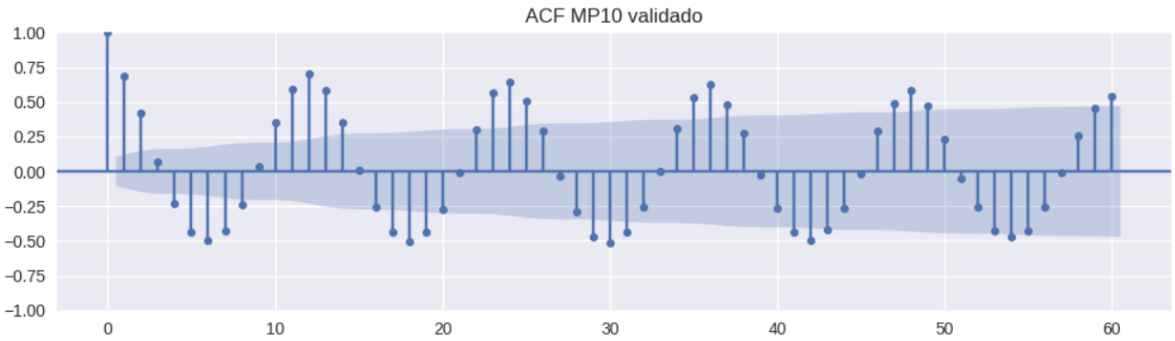

Figura 2. Gráfica de funcion de autocorrelación de los datos.

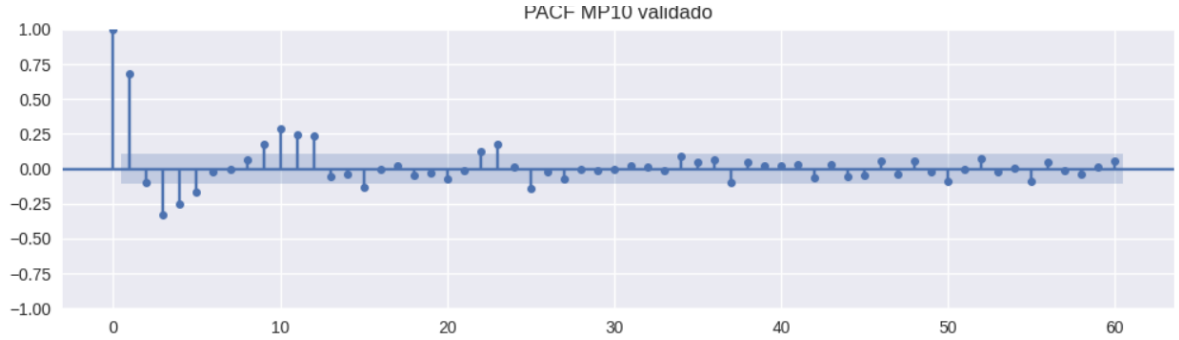

Figura 3. Gráfica de Función de Autocorrelación Parcial de los datos.

Para estudiar la dependencia temporal de la serie y orientar la elección de los órdenes del modelo, se calcularon la función de autocorrelación (ACF) y la función de autocorrelación parcial (PACF) de la serie en niveles. A continuación, se resumen de los hallazgos más relevantes:

- **Parte no estacional (rezagos cortos 1–4):**
  - La ACF presenta un valor alto en el rezago 1 (≈ 0.68) y luego decrece (0.42 en el rezago 2, 0.07 en el 3 y se vuelve negativa desde el rezago 4).
  - La PACF muestra un pico muy marcado en el rezago 1 (≈ 0.69) y luego valores mucho más pequeños, algunos negativos.

Este patrón es muy compatible con la presencia de un componente AR(1) fuerte en la parte no estacional (p = 1). En principio, también podría considerarse un modelo ARMA(1,1), pero el hecho de que la PACF se “corte” de manera clara en el primer rezago hace que un modelo con p = 1 sea una opción natural y parsimoniosa.

- **Componente estacional (rezagos cercanos a 12):**
  - En la ACF se observa un “bulto” muy evidente en torno a los rezagos 10–14, con valores altos:

        - lag 10 ≈ 0.35, lag 11 ≈ 0.59, lag 12 ≈ 0.71, lag 13 ≈ 0.58, lag 14 ≈ 0.35.
    - Esto indica una estacionalidad anual fuerte con periodo 12 (s = 12).
    - En la PACF, los rezagos cercanos a 12 muestran valores positivos moderados (~0.25–0.30), lo que sugiere la presencia de un componente estacional de bajo orden.

Un gran pico en la ACF alrededor del rezago 12, acompañado de valores moderados en la PACF, es consistente con la presencia de un término estacional de tipo AR(1) o MA(1). Por esta razón, en la parte estacional se decidió probar tanto modelos con P = 1, Q = 0 como con P = 0, Q = 1 (siempre con D = 1 y s = 12), dejando que los criterios de ajuste (AIC) y error de predicción (RMSE en el conjunto de prueba) determinaran cuál estructura se adapta mejor a los datos.

En síntesis, el análisis conjunto de la ACF y la PACF respalda el uso de un término AR(1) en la parte no estacional y un componente estacional anual de bajo orden, lo que se refleja posteriormente en la elección de modelos SARIMA con estructura (1,0,·)(·,1,·)[12], entre ellos el modelo SARIMA(1,0,1)(1,1,0)[12] seleccionado como especificación final.

3.1.3. Evaluación de estacionariedad

La teoría de los modelos ARIMA/SARIMA exige que la serie (tras las transformaciones y diferencias necesarias) sea estacionaria. Para evaluar este supuesto en la componente no estacional, se aplicó el test de Dickey-Fuller aumentado (ADF) sobre la serie mensual en niveles.

El estadístico ADF obtenido fue aproximadamente –4.03, valor que permite rechazar la hipótesis nula de raíz unitaria a niveles de significancia habituales. Esto indica que, una vez considerada la componente estacional de la serie, no es necesario aplicar una diferencia no estacional adicional (es decir, se puede trabajar con d = 0 en la parte no estacional del modelo). La estacionalidad, en cambio, se maneja mediante una diferencia estacional de orden 1 y periodo 12 (D = 1, s = 12), coherente con la estructura observada en la ACF.

Figura 4. Serie Temporal Mensual de MP10 en Parque O'Higgins.

Figura 5. Desviación Estándar Móvil de 12 Meses del MP10.

Adicionalmente, no se consideró necesario aplicar una transformación logarítmica a la serie. La desviación estándar móvil con ventana de 12 meses se mantiene, en general, en un rango aproximado entre 15 y 30 µg/m³N, con algunos periodos de mayor variabilidad al inicio de la muestra, pero sin cambios explosivos ni incrementos sostenidos en el tiempo. En este contexto, no se observa una heterocedasticidad extrema a nivel descriptivo, por lo que trabajar en niveles resulta razonable y, además, facilita la interpretación de los resultados en las unidades originales (µg/m³N).

3.1.4. Modelos SARIMA considerados a partir del análisis exploratorio

A partir de la evidencia empírica obtenida en esta sección, se puede acotar el espacio de búsqueda a una familia de modelos SARIMA relativamente parsimoniosa y coherente con la estructura observada:

- **Parte no estacional:**
  - La prueba ADF entregó evidencia de estacionariedad en la componente no estacional, por lo que se adopta **d = 0**.
  - La ACF muestra una autocorrelación muy alta en el rezago 1 (≈ 0.68) que decae en los rezagos siguientes, y la PACF presenta un pico dominante en el rezago 1 (~0.69) y luego valores mucho menores. Esto es consistente con la presencia de un término **AR(1)** fuerte, por lo que se fija **p = 1**.
  - Para capturar posibles correcciones adicionales en la estructura de corto plazo, se considera opcionalmente un término MA de bajo orden, es decir, **q = 0 o q = 1**.

- **Parte estacional (periodo s = 12):**
  - La ACF exhibe un "bulto" importante en los rezagos 10–14, con una autocorrelación muy alta en el rezago 12 (≈ 0.71), lo que indica una **estacionalidad anual marcada**.
  - Para tratar esta estacionalidad se utiliza una diferencia estacional de orden 1, es decir, **D = 1** con periodo **s = 12**.

- Los patrones de ACF y PACF en torno al rezago 12 son compatibles con un componente estacional de **bajo orden**, por lo que se consideran modelos con:
  - **P = 1 o P = 0**,
  - **Q = 0 o Q = 1**.

En resumen, el análisis exploratorio sugiere trabajar con una familia de modelos del tipo: **SARIMA(p, d, q) (P, D, Q)[12].**

Donde los órdenes considerados son:

- **Parte no estacional:**
  - p = 1
  - d = 0
  - q ∈ {0, 1}

- **Parte estacional (periodo s = 12):**
  - D = 1
  - P ∈ {0, 1}
  - Q ∈ {0, 1}

En las secciones siguientes se estiman distintas combinaciones dentro de esta familia (por ejemplo, SARIMA(1,0,0)(1,1,0)[12], SARIMA(1,0,1)(1,1,0)[12], SARIMA(1,0,0)(0,1,1)[12], entre otras) y se comparan sus desempeños utilizando el AIC y el RMSE en un conjunto de prueba, con el fin de seleccionar el modelo más adecuado.

3.2. Aplicación de técnicas y selección de modelo

Para modelar la serie de concentraciones de MP10 en la estación Parque O'Higgins, se evaluó la estacionariedad mediante el test de Dickey-Fuller aumentado (ADF), indicando la necesidad de diferenciación estacional (D=1). Basado en la estructura de autocorrelación (ACF) y autocorrelación parcial (PACF), se propusieron modelos SARIMA con componente estacional de periodo 12.

3.2.1. Criterios de Selección y Comparación de Modelos

Se compararon múltiples especificaciones manuales frente a un modelo generado automáticamente por el algoritmo *auto_arima*. La selección se basó en la minimización del Criterio de Información de Akaike (AIC) para el ajuste y la minimización de la Raíz del Error Cuadrático Medio (RMSE) evaluado en un conjunto de prueba (enero 2021 - julio 2025). Para más detalles sobre los cálculos e implementación de los distintos modelos, revisar el Jupyter Notebook del repositorio que acompaña este documento.

A continuación, se presenta la tabla comparativa con los resultados de los modelos más relevantes:

Tabla 1. Comparación de desempeño entre modelos candidatos

Modelo	Especificación	AIC (Entrenamiento)	RMSE (Prueba)	Decisión
Propuesto (Manual)	**SARIMA(1,0,1)(1,1,0)[12]**	**2051.3**	**10.03**	**Seleccionado**
Auto ARIMA	SARIMA(1,0,1)(0,1,1)[12]	2109.6	11.32	Descartado
Alternativa 1	SARIMA(1,0,0)(1,1,0)[12]	2063.7	10.25	Descartado
Alternativa 2	SARIMA(0,0,0)(1,1,0)[12]	2091.6	10.27	Descartado

Fuente: Elaboración propia basada en datos validados de SINCA.

Como se observa en la Tabla 1, el modelo seleccionado SARIMA(1,0,1)(1,1,0)[12] (sin tendencia determinista), el cual superó al modelo automático tanto en ajuste (menor AIC) como en capacidad predictiva (menor RMSE), reduciendo el error de pronóstico en aproximadamente un 11% respecto a la referencia automática.

3.2.2. Diagnóstico de Idoneidad

El análisis de los residuos del modelo seleccionado muestra que estos se distribuyen en torno a cero con una desviación estándar estable, aunque persisten algunos valores atípicos (outliers) asociados al año 1997.

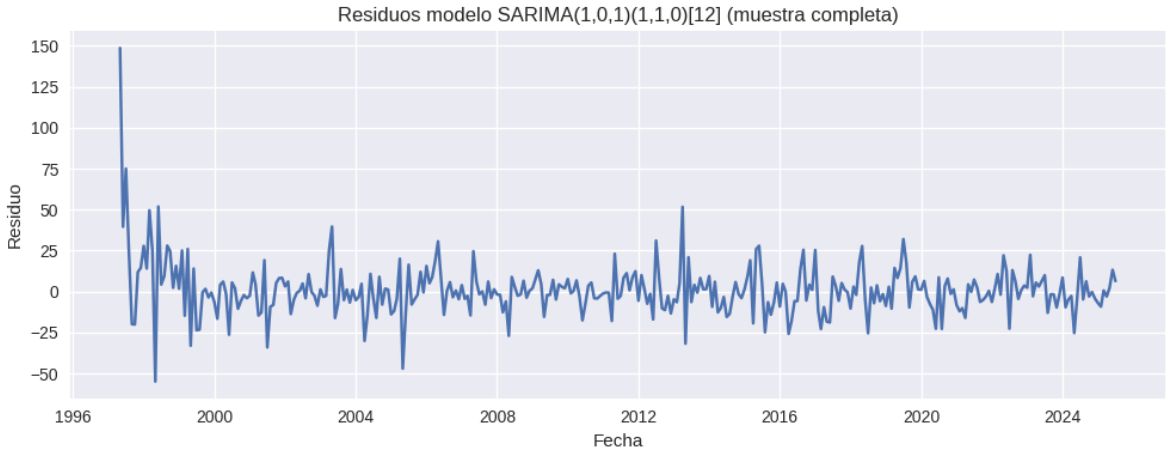

Figura 6. Función de Autocorrelación (ACF) de los residuos del modelo seleccionado.

La prueba de Ljung-Box indica que existe autocorrelación remanente en los rezagos estacionales, lo cual es una limitación común en series de contaminación ambiental

con alta variabilidad exógena. Sin embargo, el modelo se considera apropiado para el objetivo del estudio dado su alto rendimiento predictivo validado.

Para explorar si era posible mejorar el ajuste, se probó incluir una variable dummy asociada a los primeros meses de la serie (Modelo con dummy: SARIMA(1,0,1)(1,1,0)[12] + dummy_1997 como regresor exógeno). El modelo con dummy arrojó parámetros similares para las componentes AR y MA, y la dummy no resultó estadísticamente relevante. Además:

- El AIC del modelo con dummy fue ligeramente mayor (≈ 2053.3) que el del modelo sin dummy (≈ 2051.3) cuando se estima sobre el conjunto de entrenamiento.
- El RMSE en el conjunto de prueba fue prácticamente igual (≈ 10.02 µg/m³N), sin mejora sustantiva respecto al modelo base (≈ 10.03 µg/m³N).

Dado que la dummy no mejora de forma relevante ni el criterio de información ni el desempeño predictivo, y añade complejidad a la especificación, se decidió mantener el modelo SARIMA(1,0,1)(1,1,0)[12] sin regresores adicionales.

3.3. Predicciones e intervalos de confianza

Utilizando el modelo SARIMA(1,0,1)(1,1,0)[12], reentrenado con la totalidad de la muestra validada (1997-2025), se generaron predicciones para un horizonte de 12 meses.

3.3.1. Proyección a Corto Plazo

La siguiente tabla detalla los valores esperados para los primeros tres meses del horizonte de predicción, junto con sus intervalos de confianza al 95%:

Tabla 2. Pronóstico de MP10 (agosto - octubre 2025)

Fecha	Predicción (µg/m ^ 3N)	Límite Inferior (95%)	Límite Superior (95%)
Agosto 2025	**70.9**	**47.3**	**94.4**
Septiembre 2025	**60.0**	**36.2**	**83.9**
Octubre 2025	**49.0**	**24.9**	**73.1**

Fuente: Elaboración propia

3.3.2. Validación Externa

Al comparar estas cifras con los datos preliminares reportados recientemente por SINCA, el modelo presentó un error promedio (RMSE) de solo **3.7 µ g/m^3N** confirmando su alta capacidad para anticipar la dinámica actual de la serie.

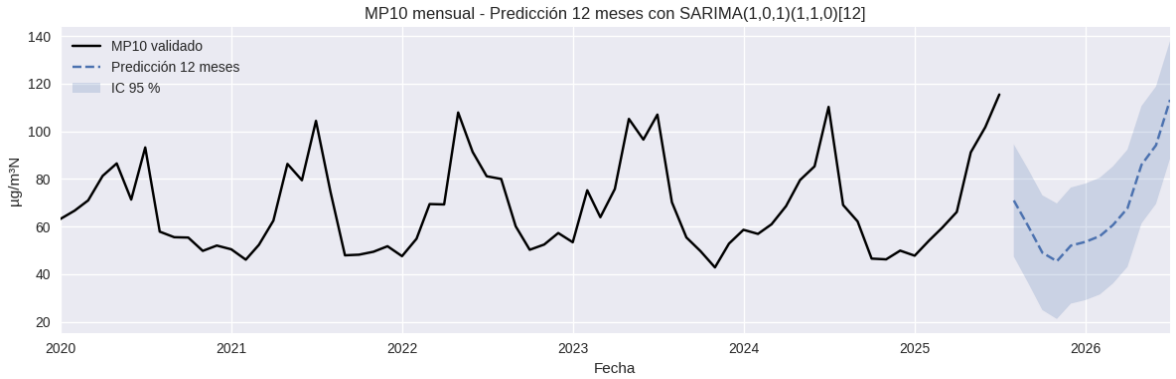

Figura 7: Serie histórica reciente (desde 2020) y pronóstico a 12 meses con intervalos de confianza.

Como se aprecia en la Figura 7, la proyección captura correctamente la estacionalidad anual, anticipando el descenso de las concentraciones hacia la primavera y verano, comportamiento coherente con la historia de la estación Parque O'Higgins.

4. CONCLUSIONES

A continuación, se presentan las conclusiones del trabajo realizado y los resultados obtenidos expuestos en las secciones anteriores.

En primer lugar, el análisis exploratorio mostró una marcada dependencia temporal y una clara componente estacional anual, coherente con la dinámica de la contaminación por material particulado en la ciudad. El test de Dickey–Fuller aumentado entregó evidencia de estacionariedad en la componente no estacional, por lo que se adoptó una especificación con diferencia estacional de orden 1 y periodo 12, manteniendo la serie en niveles. Esta decisión permitió trabajar directamente con las concentraciones de MP10 en unidades de µg/m³N, lo que facilita la interpretación de los resultados en el contexto ambiental.

A partir del análisis de la ACF y la PACF, y de la comparación sistemática de distintos candidatos, se seleccionó el modelo SARIMA(1,0,1)(1,1,0)[12]. Este modelo obtuvo un error cuadrático medio (RMSE) de alrededor de 10 µg/m³N al predecir el periodo 2021–2025, desempeño que resultó mejor que el del modelo escogido por el procedimiento automático auto_arima, tanto en términos de RMSE como de criterio de información (AIC). Esto indica que una selección guiada por el

análisis gráfico y los criterios de ajuste puede producir modelos más competitivos que una búsqueda completamente automatizada, al menos para esta serie.

El diagnóstico de residuos mostró que, si bien estos se encuentran razonablemente centrados en torno a cero y con una variabilidad compatible con la de la serie ajustada, existen errores muy grandes en los primeros meses del periodo analizado y se detecta autocorrelación residual pequeña pero estadísticamente significativa en rezagos estacionales según la prueba de Ljung–Box. Por lo tanto, el modelo no reproduce perfectamente toda la dependencia temporal ni explica completamente algunos episodios extremos al inicio de la serie. Se exploró la inclusión de una variable dummy para los primeros meses de 1997, pero no produjo mejoras relevantes en AIC ni en RMSE, por lo que se mantuvo la especificación base. Aun así, el comportamiento general de los residuos y el buen desempeño predictivo sugieren que el modelo es suficientemente adecuado para describir la dinámica global de MP10 en la estación estudiada, reconociendo explícitamente la presencia de outliers tempranos y cierta autocorrelación residual como principales limitaciones.

En cuanto a las predicciones, el modelo SARIMA(1,0,1)(1,1,0)[12] se utilizó para proyectar un horizonte de 12 meses posteriores al último dato validado. Para los meses de agosto, septiembre y octubre de 2025, el modelo proyecta concentraciones medias cercanas a 70,9; 60,0 y 49,0 µg/m³N, respectivamente, con intervalos de confianza del 95 % que abarcan, según el mes, rangos aproximados entre 25 y 95 µg/m³N. Estas predicciones se interpretan como los niveles promedio mensuales de MP10 que cabría esperar en ausencia de cambios estructurales importantes en las condiciones de emisión o meteorología, y mantienen el patrón estacional observado históricamente, con niveles más altos en los meses fríos y una reducción posterior.

Finalmente, al comparar estas predicciones con los valores preliminares de MP10 reportados por SINCA para agosto, septiembre y octubre de 2025, se obtuvo un RMSE cercano a 3,7 µg/m³N, considerablemente menor que la variabilidad típica de la serie y que la desviación estándar de los residuos del modelo. Aunque se trata de una comparación basada en muy pocos datos (N = 3) y sobre valores aún sujetos a validación, este resultado sugiere que el modelo es capaz de reproducir de manera razonable el comportamiento reciente de la serie y ofrece predicciones útiles a muy corto plazo. En conjunto, el trabajo cumple el objetivo de modelar y predecir la evolución de MP10 en Parque O’Higgins y proporciona una primera herramienta para anticipar la evolución de las concentraciones mensuales, que puede servir de base para futuros análisis que incorporen variables explicativas adicionales o modelos más complejos (por ejemplo, con tratamiento específico de outliers o heterocedasticidad).

5. REFERENCIAS

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control. Wiley.

Gobierno de Chile, Ministerio del Medio Ambiente. (s. f.). Región Metropolitana de Santiago – Estaciones de monitoreo de la calidad del aire. Sistema de Información Nacional de Calidad del Aire. https://sinca.mma.gob.cl/index.php/region/index/id/M