"Análisis de la Relación entre la Contaminación Atmosférica y las Condiciones Meteorológicas: Un Enfoque PCA y PLS"

Autor:
Rytck Iuliia

Valencia 2023

Índice

Resumen

Este trabajo se enfoca en el análisis de la calidad del aire en la ciudad de Valencia
Se lleva a cabo un análisis de correlación entre la concentración de partículas en el aire y las condiciones meteorológicas. Se busca comprender cómo las diferentes condiciones meteorológicas pueden afectar la calidad del aire. Este análisis proporciona información sobre las relaciones y dependencias entre las variables, lo que ayuda a identificar los factores meteorológicos más influyentes.
Con base en los resultados obtenidos, se desarrolla un modelo predictivo para predecir la concentración de partículas en el aire en función de las condiciones meteorológicas y la evolución de la concentración de partículas a lo largo del tiempo. Este modelo brinda una herramienta útil para anticipar niveles de contaminación y tomar medidas preventivas para mejorar la calidad del aire en Valencia.
En conclusión, este trabajo destaca la importancia de comprender la relación entre la calidad del aire, las condiciones meteorológicas y su evolución temporal. Los resultados obtenidos proporcionan información valiosa para la toma de decisiones y la implementación de estrategias para mejorar la calidad del aire en la ciudad. La combinación de técnicas de análisis exploratorio, correlación y modelado predictivo ofrece una perspectiva integral que invita a profundizar en el análisis y comprensión de la calidad del aire en Valencia.

Introducción

El presente trabajo se centra en el análisis de la calidad del aire en la ciudad de Valencia utilizando la base de datos del Ayuntamiento de Valencia, conocida como el Sistema de Información Territorial de Valencia (SIT-Valencia). Esta base de datos proporciona información valiosa sobre la concentración de partículas en el aire y las condiciones meteorológicas en la ciudad.

La motivación detrás de este trabajo radica en la importancia de comprender y evaluar la calidad del aire en un entorno urbano. La contaminación del aire es un problema global que afecta la salud humana y el medio ambiente, y es crucial para las autoridades contar con herramientas efectivas para monitorear y abordar esta problemática.

En este sentido, se han aplicado diversas técnicas de análisis de datos para explorar y comprender la relación entre las variables de concentración de partículas en el aire y las condiciones meteorológicas en Valencia. Se ha realizado un análisis exploratorio mediante el uso de Análisis de Componentes Principales (PCA) para identificar patrones y estructuras en los datos. Además, se ha utilizado el método de Regresión por Mínimos Cuadrados Parciales (PLS) para examinar la relación entre las variables explicativas (X) y las variables respuesta (Y), lo que nos permite evaluar la influencia de las condiciones meteorológicas en la calidad del aire.

Para tener en cuenta la dependencia temporal en los datos y comprender la evolución de la relación entre las variables, se ha aplicado el modelo PLS dinámico.

Finalmente, se ha llevado a cabo una predicción de la variable NO2 utilizando los modelos desarrollados, lo que nos brinda una herramienta para anticipar la concentración de partículas en el aire y tomar medidas preventivas.

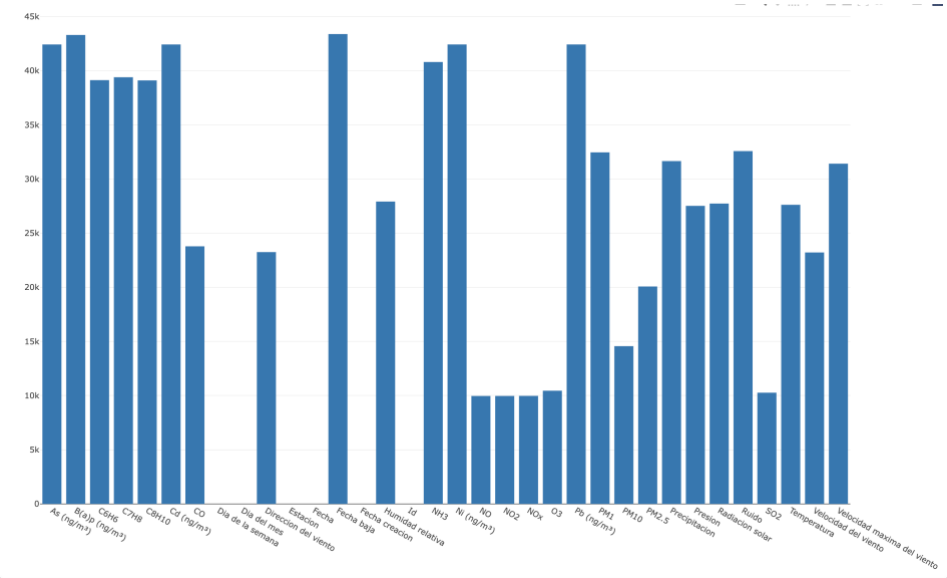
Material y Métodos

La base de datos del Ayuntamiento de Valencia, conocida como el Sistema de Información Territorial de Valencia (SIT-Valencia), proporciona información sobre la calidad del aire en la ciudad. Esta base de datos es una herramienta de gestión de información geográfica que recopila y almacena datos sobre la ciudad.

La base de datos consta de 43,388 observaciones y 34 columnas. De estas, 19 variables cuantitativas representan la concentración de partículas en el aire, incluyendo PM1, PM2.5, PM10, NO, NO2, NOx, O3, SO2, CO, NH3, C7H8, C6H6, ruido, C8H10, As (ng/m³), Ni (ng/m³) y Cd (ng/m³). Además, se incluyen otras 8 variables cuantitativas que corresponden a las condiciones meteorológicas, como la velocidad y dirección del viento, temperatura, humedad relativa, presión, radiación solar, precipitación y velocidad máxima del viento. También se registran variables adicionales como el ID, fecha, día de la semana, día del mes y estación.

La estructura de la base de datos se organiza en 13 bloques que corresponden a las 13 estaciones meteorológicas. Cada estación ha sido observada en varios momentos del tiempo desde el 1 de enero de 2004 hasta el 31 de diciembre de 2022.

El primer problema que aparece son los datos faltantes. Entre las 34 variables 28 tienen valores NA. En la siguiente figura están resumidos número de NA´s por variable.



Para este trabajo se ha decidido eliminar las variables con más del 80%, de NA´s, que son: As, B, Pb, Cd, Ni, C6H6, C7H8, C8H10, NH3.

Para el resto de variables se ha decidido reducir la frecuencia de los datos a media semanal. Tras eliminación y agrupación por semana se obtiene un dataframe dimensión 6247x20. Se ha mejorado el porcentaje de los datos faltantes, pero aún hay variables con los datos faltantes. A continuación, se ha empleado las técnicas de imputación con el paquete mice de R.

Después de tratar los datos faltantes aparece otro problema: Los valores en cada estación no están observadas en el mismo rango de tiempo entre 2004-2022. La siguiente figura muestra el rango de observaciones para cada estación.

Estacion <chr>	fecha_min <chr>	fecha_max <chr>
Avda. Francia	2009 01	2022 52
Bulevard Sud	2010 01	2022 52
Conselleria Meteo	2012 01	2022 52
Moli del Sol	2009 01	2022 52
Nazaret Meteo	2020 01	2022 52
Pista Silla	2004 01	2022 52
Politecnico	2008 01	2022 52
Puerto Moll Trans. Ponent	2021 01	2022 52
Puerto Valencia	2017 01	2018 52
Puerto llit antic Turia	2021 01	2022 52
Valencia Centro	2018 01	2022 52
Valencia Olivereta	2022 01	2022 52
Viveros	2004 01	2022 52

Para el objetivo de este trabajo se ha decidido seleccionar las estaciones que tendrán observaciones para los últimos 10 años, desde 01/2012 hasta 12/2022.

El conjunto de datos final consta de 3991 observaciones, agrupado en 7 bloques, que se refieren a las 7 estaciones y 20 variables:

Las variables respuesta, las que se pretende estudiar: PM1; PM2.5; PM10 - son partículas suspendidas en el aire de diferentes tamaños que pueden afectar la salud humana.

NO; NO2; NOx - óxidos de nitrógeno que se emiten principalmente por la quema de combustibles fósiles.

O3 - el ozono troposférico, un gas que se forma cuando los óxidos de nitrógeno y los compuestos orgánicos volátiles reaccionan con la luz solar.

SO2 - el dióxido de azufre, un gas que se produce principalmente por la quema de combustibles fósiles.

CO - el monóxido de carbono, un gas inodoro e incoloro que se produce por la quema incompleta de combustibles fósiles.

Y las variables explicativas, que se refieren a las condiciones meteorológicas:

Viento (Velocidad y Dirección) - influye en la dispersión de los contaminantes en el aire y puede afectar la calidad del aire en áreas cercanas a fuentes de emisión.

Temperatura: influye en la formación de ozono y en la reactividad de otros contaminantes en el aire. La temperatura también puede afectar la tasa de emisión de contaminantes.

Humedad: influye en la formación y el transporte de contaminantes en el aire. La humedad también puede afectar la deposición de partículas en superficies.

Presión: influye en la formación y el transporte de contaminantes en el aire, así como en la dispersión de los contaminantes.

Radiación solar: influye en la formación de ozono y en la reactividad de otros contaminantes en el aire. La radiación solar también puede afectar la temperatura y la tasa de emisión de contaminantes.

Precipitación: puede limpiar el aire de los contaminantes y reducir los niveles de partículas en suspensión y gases contaminantes.

Para el análisis exploratorio y estudio de relaciones entre las variables se hace uso de la técnica no supervisada Análisis de Componentes Principales.

Con el objetivo de estudiar las relaciones entre las variables respuesta y variables explicativas se emplea la técnica supervisada de PLS y PLS dinámico con decalaje en 5 intervalos. Además, se construyen modelos con el objetivo de predecir las emisiones de óxidos de nitrógeno.

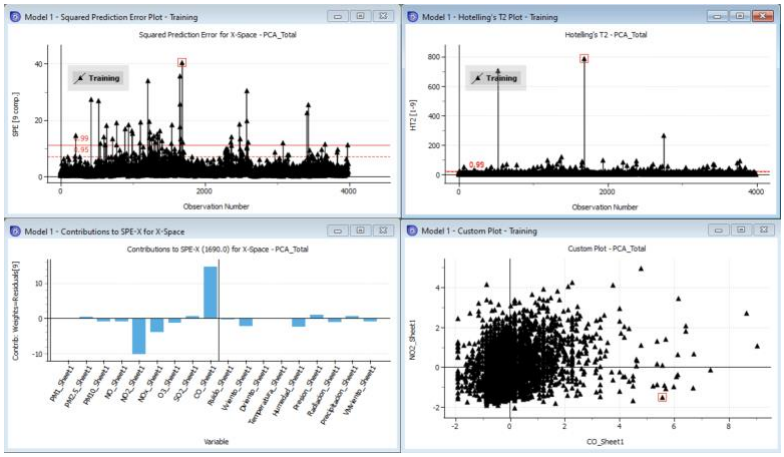
Como análisis adicional se emplea la técnica de análisis de datos funcionales para analizar la evolución del contenido de O3 para las 7 estaciones.

Resultados

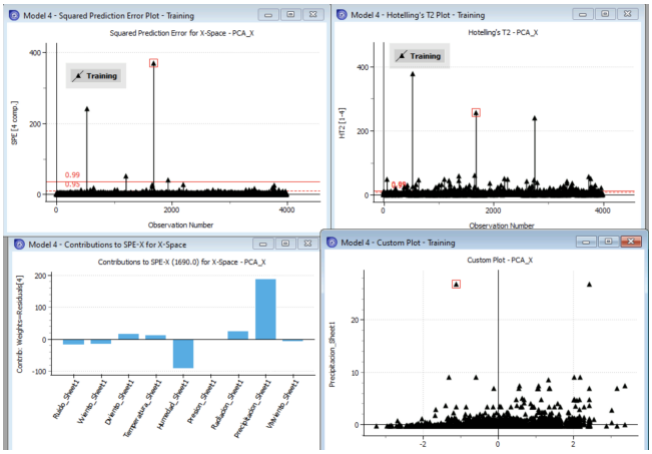
Modelo PCA

Como primer paso se ajusta el modelo PCA con todas las variables (bloque X y bloque Y). El modelo ajustado extrae 3 componentes principales con el porcentaje de variabilidad explicada de R^2=54% y capacidad predictiva Q^2=0,53.

Para asegurarnos de que no hay problemas en los datos se realiza el diagnostico a través del análisis de los residuos con el gráfico SPE y las observaciones en el espacio de las 3 componentes con el gráfico de T^2 de Hotelling. Observando el gráfico SPE vemos que aparecen valores anómalos, que están por encima del límite de 0.99. Es esperable que 1% de las observaciones van a estar fuera de ese límite, que son aproximadamente 40. Al observar el gráfico de contribución se ve que la variable CO está muy por encima de la media y la variable NO2 está muy por debajo de la media. La misma observación aparece como extrema en el gráfico de T^2 porque tiene un valor demasiado alto. Para determinar si se ha roto la estructura de correlación, se representan estas dos variables en un gráfico y se puede ver que no existe correlación y por tanto no rompe la estructura. Por lo que se ha decidido no eliminarlo.



Aparecen datos anómalos que se alejan de la media y rompen con la estructura de correlación. En concreto la relación entre precipitación y humedad. Para preservar la integridad y consistencia de la estructura de correlación se procede a eliminar las observaciones anómalas (nº1690 y nº531)



Se procede a analizar el modelo.

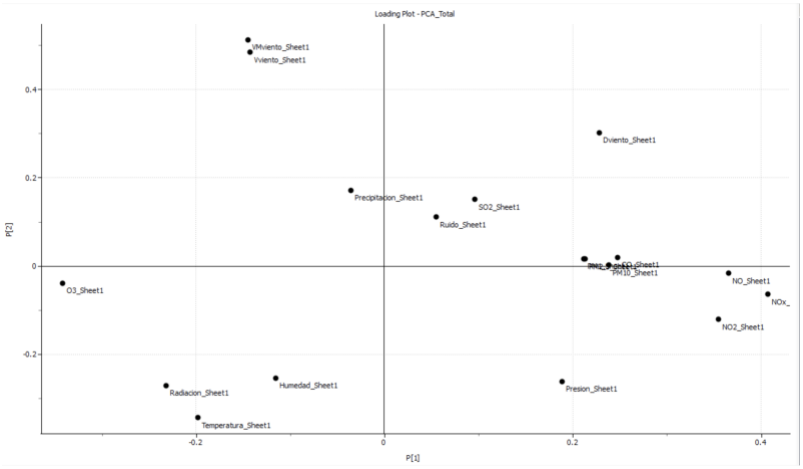
Primero se analiza el grafico de loadings que muestra la contribución de cada variable en el espacio de las primeras dos componentes.

En el gráfico de loadings analizando la dirección y magnitud de los vectores de cada variable. Las variables de óxidos de nitrógeno (NO´s) y el ozono (O3) tienen una correlación alta y una mayor influencia en la formación de la primera componente principal. En cuanto a la segunda componente las variables que más influencia tienen son la velocidad del viento (Vviento, VMviento) y Radiación, Humedad y Temperatura.

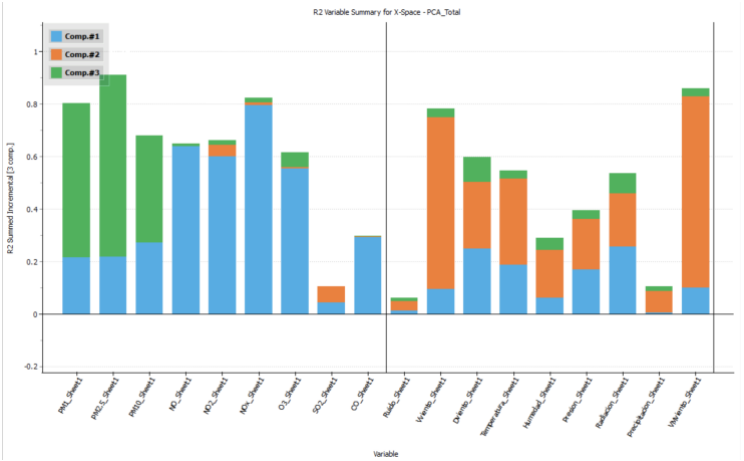
Además, las variables de PM (partículas suspendidas) y monóxido de carbono (CO) que se encuentran cerca lo que indica que están correlacionadas entre sí. Por lo tanto, podrían representar una característica subyacente del conjunto de datos.

En resumen, el gráfico de loadings en el modelo PCA es una herramienta visual que permite interpretar la importancia y relación de las variables originales en cada

componente principal, lo que puede ayudar a identificar patrones y estructuras en los datos.



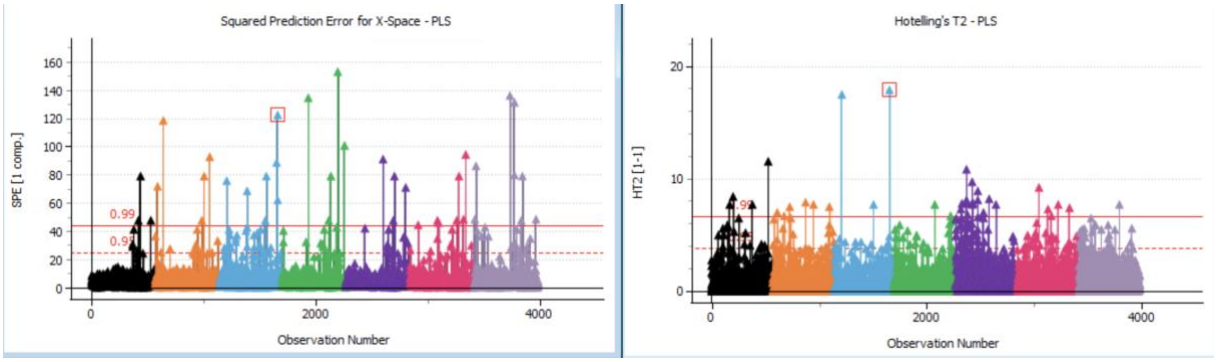
Al analizar el resumen de R^2 de cada componente para cada variable se observa que la primera componente esta fundamentalmente explicada por las variables de óxidos de nitrógeno (NO, NOx, NO2) y ozono (O3). En menor medida por CO, y las partículas suspendidas (PM´s), que tiene más peso en la componente 2. La R^2 de la variable SO2 es muy baja repartida entre las dos componentes. Y la componente 3 está explicada en mayor parte por las variables de las condiciones meteorológicas, fundamentalmente por Velocidad de viento. El Ruido y Precipitación no tienen casi peso en ninguna componente.



Modelo PLS

A continuación, se realiza el aprendizaje supervisado ajustando modelos PLS y PLS dinámico.

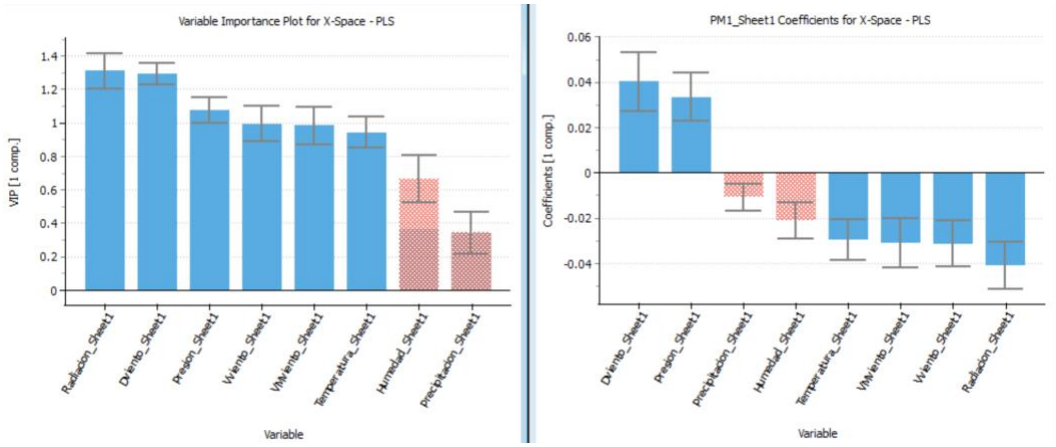
Primero se ajusta el modelo PLS con todas las variables excluyendo la variable Ruido ya que no es característico de condiciones meteorológicas. El diagnóstico indica que no hay datos anómalos, que rompen con la estructura de correlación.



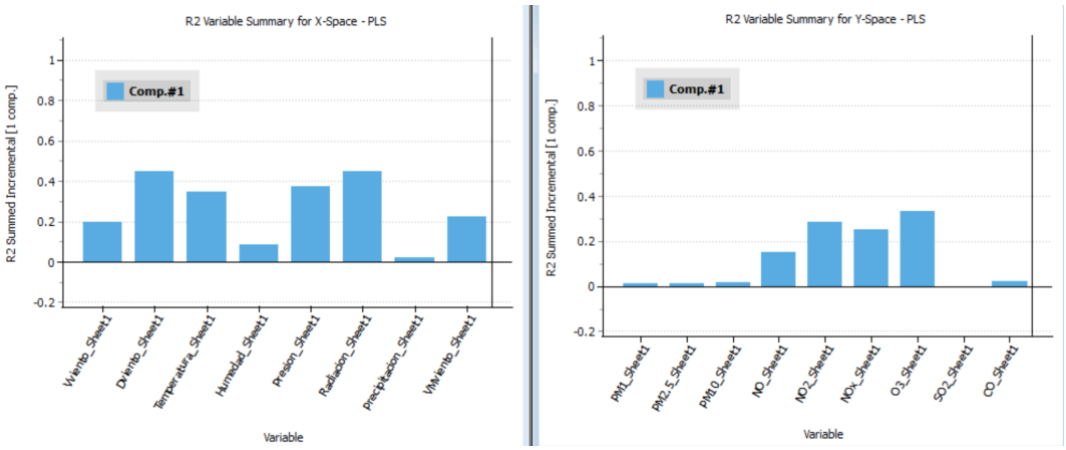
Vamos a conectar las variables características de calidad de aire con las variables meteorológicas para encontrar el subespacio X y subespacio Y de manera que las variables latentes que se genera el modelo tengan la máxima covarianza.

El primer modelo PLS se construye con 1 componente y R2 muy baja de 0,12 capacidad predictiva Q2 de 0,12.

Utilizando VIP y coeficientes de regresión se decide eliminar la variable Precipitación y dejar la variable Humedad, ya que tiene un coeficiente de regresión considerable.

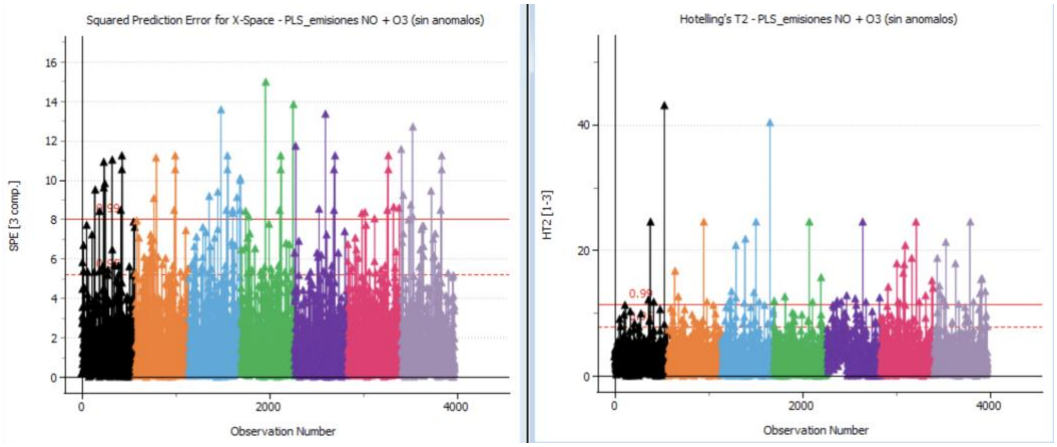


Viendo el resumen de R^2, las variables PM1; PM2.5; PM10; SO2 y CO no aportan casi nada a la componente 1 por lo que no guardan relación con el espacio X. Igual que la Precipitación en el espacio X.



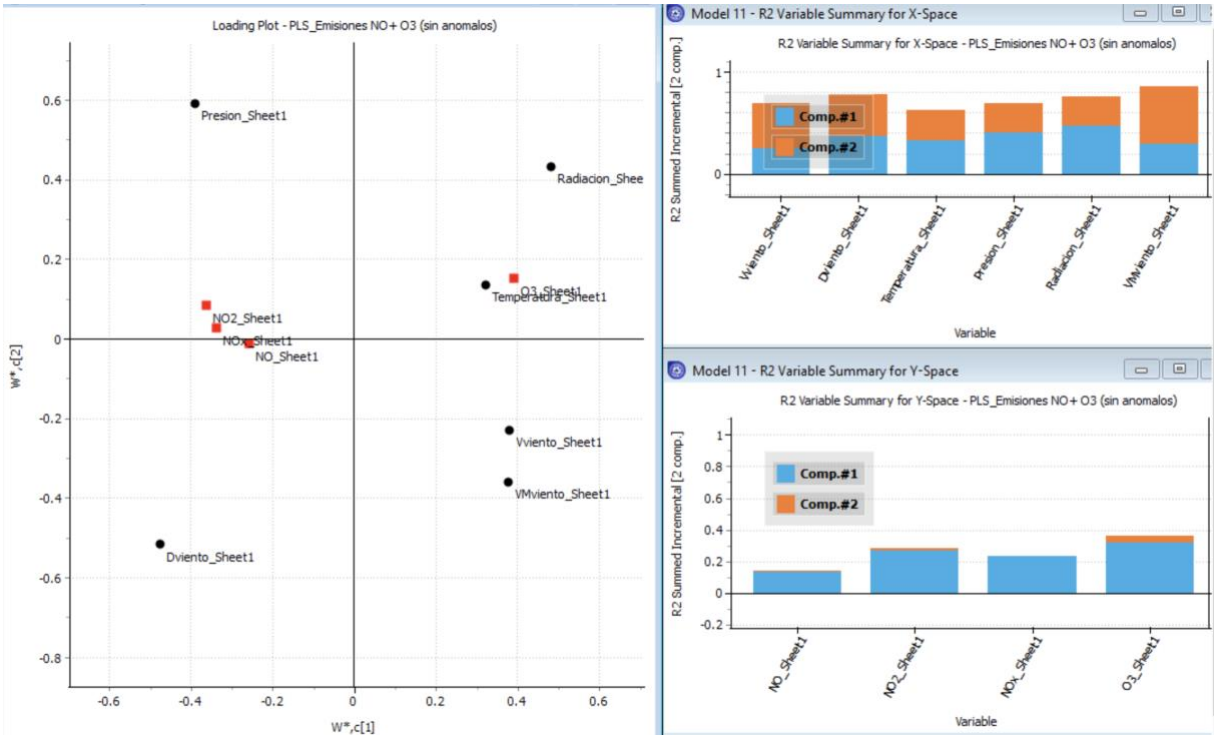
Se ajusta el modelo PLS con variables Y(NO, NO2, NOx y O3) del bloque emisiones y con X prescindiendo de Humedad, Precipitación.

Diagnostico. Análisis de los residuos (SPE - Suma Cuadrados Residual) aparecen algunas observaciones por encima del límite de 0.99, pero no supera el doble del intervalo. Por lo tanto, podríamos decir que tenemos un conjunto de datos sin observaciones atípicas y extremas.



Pasamos a interpretar el modelo. Las dos primeras componentes explican 25,8% de la variabilidad. En primer lugar, vamos a analizar la relación interna con el gráfico T1-U1. La relación entre los scores T y U. Se observa que para componente 1 hay una relación lineal. Conforme añadimos los componentes esa relación se va debilitando.

Analizamos la R^2 del espacio X y espacio Y junto con el gráfico de loadings y pesos para X y Y (W*C1-W*c2). En los siguientes gráficos observamos la R2 del espacio X y R2 del espacio Y, que nos indica como estos componentes explican las variables respuesta y los regresores.



En Cuanto a las variables Respuesta, se observa que NO2, NO, NOx y O3 están explicadas fundamentalmente por la componente 1. La componente 2 no aporta casi nada a la R2. En cuanto a las variables X, están explicadas por las dos componentes en más o menos misma proporción.

Lo mismo podemos observar en el gráfico de loadings. Los puntos rojos-las variables respuesta están fundamentalmente explicadas por la componente 1. Las variables NO2, NOx; NO y O3 tienen muy poco peso en la componente 2. La O3 a su vez está negativamente relacionada con las tres primeras.

En cuanto a los regresores en el gráfico de loadings, Presión tiene más peso en la componente 2 y la variable Temperatura tiene más peso en la componente 1 y es la que mayor correlación tiene con los regresores. Temperatura esta positivamente relacionada con O3 y tiene correlación negativa con NO´s.

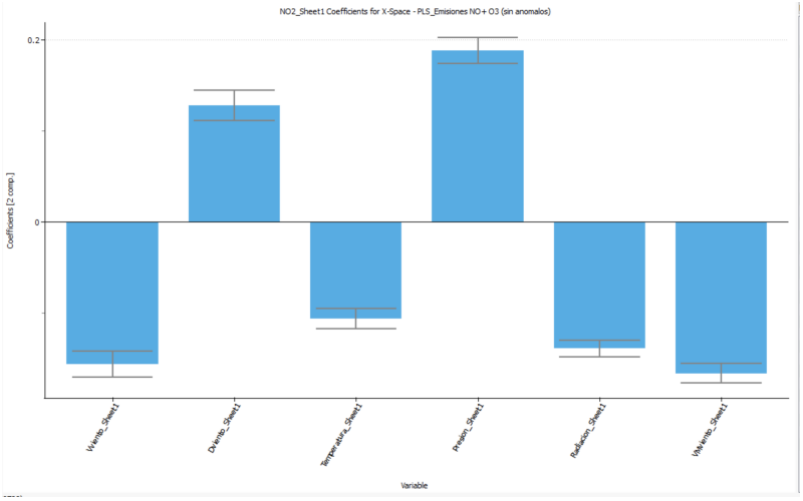
También podemos observar en el gráfico de loadings dos direcciones de variabilidad. Una que está asociada a niveles de Radiación y esta positivamente relacionada con Temperatura y niveles de O3. Lo que tiene sentido porque cuanto más intensa sea la radiación solar, mayor será la cantidad de calor absorbido y, por lo tanto, más alta será la temperatura. Y según expertos altos niveles de radiación UV contribuyen a la formación de ozono troposférico en áreas urbanas y regiones con altos niveles de contaminantes. Se forma a través de reacciones químicas complejas entre contaminantes atmosféricos, como los óxidos de nitrógeno (NOx) y los compuestos orgánicos volátiles (COV), en presencia de luz solar (radiación) y altas temperaturas. La temperatura también puede influir en los niveles de ozono. En general, las reacciones químicas que producen ozono son más rápidas a temperaturas más altas. Por lo tanto, en climas cálidos, especialmente en verano, es más probable que se formen niveles más altos de ozono troposférico. La dirección del viento tiene un impacto en cuanto al transporte de contaminantes desde

áreas de origen hacia otras regiones. La dirección del viento también puede afectar la cantidad de radiación solar que llega a una determinada ubicación y la temperatura resultante. Si el viento proviene de una región con nubes o contaminantes que bloquean la radiación solar, puede haber una disminución en la cantidad de radiación y, por lo tanto, una menor contribución a la temperatura y la formación de ozono.

Es importante tener en cuenta que la relación entre la dirección del viento, la radiación, la temperatura y los niveles de O3 puede variar en diferentes situaciones y ubicaciones. Los patrones atmosféricos, la topografía local y otras condiciones pueden influir en cómo se relacionan estos factores específicamente en un área determinada.

Otra dirección ortogonal a la primera está asociada a las condiciones meteorológicas de Nivel de Presión, Velocidad del viento, y emisiones de NO´s, pero no afecta a estas emisiones.

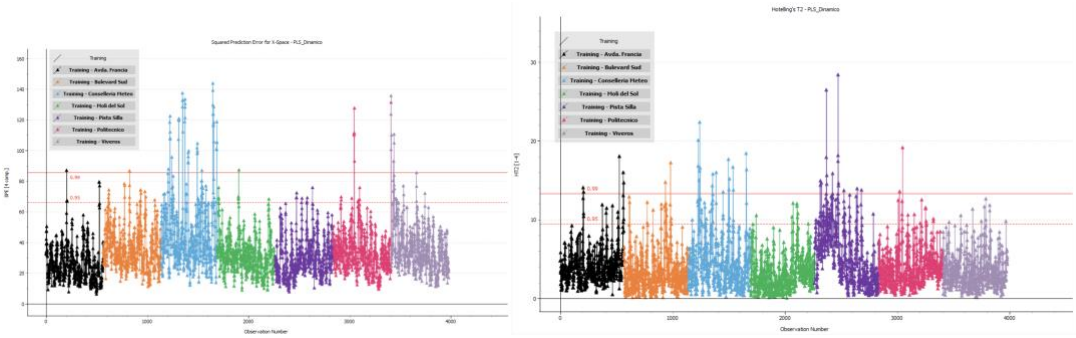
Expresamos el modelo en función de los coeficientes de regresión para la variable NO2 (Gráfico de Coeficientes de regresión para NO2). El modelo resultante tiene coeficientes positivos para la Radiación, Temperatura, y Velocidad del Viento.



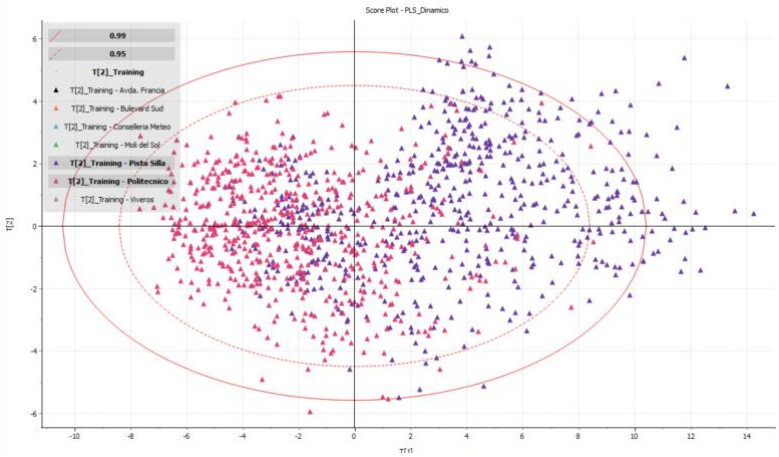
PLS Dinámico

Se ajusta un modelo PLS dinámico con 5 decalajes para capturar la dependencia temporal, mejorar la precisión de las predicciones, analizar la evolución temporal y obtener una mejor interpretación de los resultados.

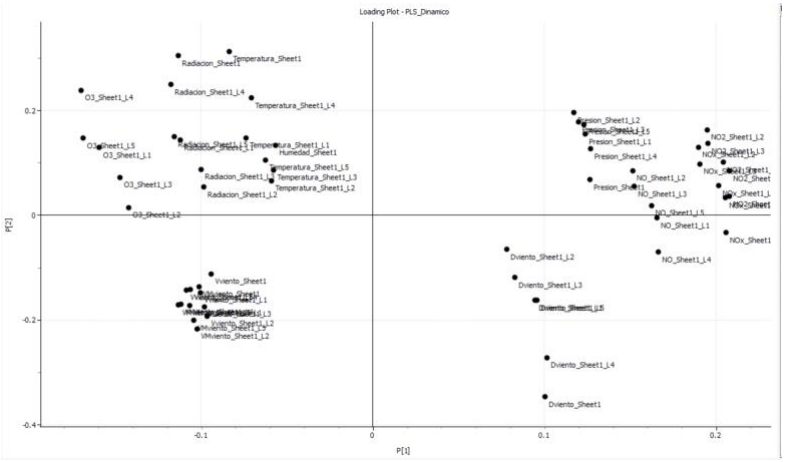
Al realizar le diagnóstico no se observan los datos anomalos, pero se observa que para la estación Pista Silla (color morado) tiene valores T2 más elevados y la Conselleria Meteo (azul) los valores SPE también por encima de 0.99



Al estudiar los gráficos de scores también observamos que la estación Pista Silla y Politécnico son las que más contribuyen a la PC2 y están negativamente relacionadas.. Politécnico tiene scores negativos en PC1 y Pista Silla positivos. La componente 2 contribuye a la separación entre las observaciones pero en menor medida.

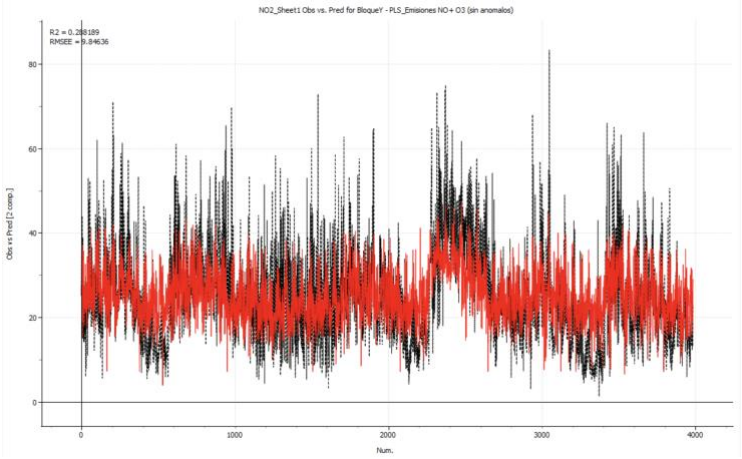


Observando los gráficos de loadings nos indica que la variabilidad se explica por las variables NO2 y NOx y O3. Por ejemplo, NO toma valores más altos en estación Pista silla y más bajos en Politécnico. Y O3 al revés, En Politécnico O3 más alto mientras en Pista Silla más bajo.

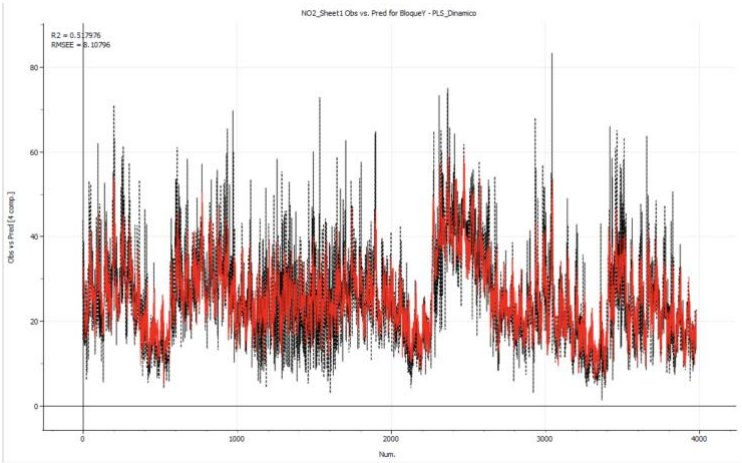


Modelo de Predicción para NO2

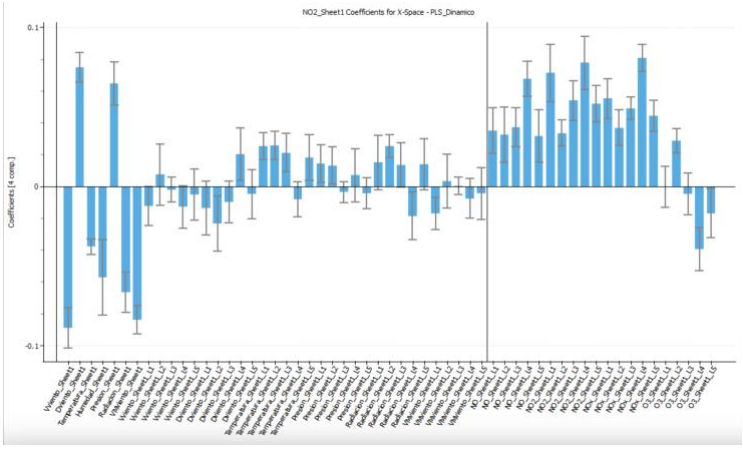
Primero se construye un modelo predictivo con PLS sin dinámica. El ajuste es bastante bajo con R2=0,288 y RMSE=9,846



Sin embargo, al crear un modelo predictivo PLS teniendo en cuenta la dinámica mejora el ajuste de predicción. R2=0,517 y RMSE = 8,10 más bajo que el modelo anterior.



Los Coeficientes de Regresión para la predicción de NO2



Conclusiones

El modelo PCA fue ajustado con todas las variables, extrayendo 3 componentes principales que explican el 54% de la variabilidad y tienen una capacidad predictiva (Q2) de 0,53. El análisis de los loadings reveló que las variables de óxidos de nitrógeno (NOx) y ozono (O3) tienen una correlación alta y una mayor influencia en la primera componente principal. Además, se identificó una correlación entre las variables de partículas suspendidas (PM) y monóxido de carbono (CO). Estos resultados sugieren la existencia de patrones y estructuras en los datos relacionados con la calidad del aire y las condiciones meteorológicas.

Se ajustaron modelos PLS y PLS dinámico para establecer la relación entre las variables de calidad del aire y las variables meteorológicas. El análisis de los loadings y los coeficientes de regresión reveló que las variables de calidad del aire, como NO2, NO, NOx y O3, están principalmente explicadas por la primera componente. Las variables meteorológicas, como la temperatura y la velocidad del viento, también tienen una influencia significativa en la explicación de las variables de calidad del aire. Estos hallazgos resaltan la importancia de las condiciones meteorológicas en la formación y concentración de contaminantes atmosféricos.

El análisis de los loadings mostró una dirección de variabilidad asociada a niveles de radiación solar, que está positivamente relacionada con la temperatura y los niveles de O3. Esto concuerda con la influencia de la radiación solar en la formación de ozono troposférico y destaca la importancia de la temperatura en las reacciones químicas que generan ozono. La dirección del viento también juega un papel crucial en el transporte de contaminantes y puede afectar la radiación solar y la temperatura en una ubicación determinada.

Se construyó un modelo predictivo utilizando PLS sin dinámica, que mostró un ajuste bajo con R2=0,288 y RMSE=9,846. Sin embargo, al considerar la dinámica en el modelo PLS, se mejoró el ajuste de predicción, obteniendo un R2=0,517 y un RMSE=8,10. Esto demuestra la importancia de tener en cuenta la dependencia temporal para una predicción más precisa de los niveles de NO2.

Este estudio resalta la relación entre la calidad del aire, las condiciones meteorológicas y las variables de contaminación atmosférica. Los resultados obtenidos proporcionan información valiosa para comprender los patrones y las influencias que afectan la calidad del aire en la ciudad de Valencia. La combinación de técnicas como el PCA y el PLS permite identificar variables clave, interpretar la relación entre las variables y desarrollar modelos predictivos útiles para la toma de decisiones y la implementación de medidas de control de la calidad del aire.