Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

Examen

Modelo de Procesos Gaussianos

Autor: Gabriel Ortega
Profesor: Nicolás Caro
Auxiliar: Rodrigo Lara

Fecha de entrega: 22 de agosto de 2020
Santiago, Chile

Resumen

El siguiente reporte corresponde al examen del curso de laboratorio de ciencia de datos. A modo de contexto, el conjunto de datos a trabajar consiste en el reporte estadístico oficial de calidad del aire en Beijing. Este conjunto esta compuesto por registros de frecuencia horaria en 11 estaciones de monitoreo ubicadas en distintos puntos de la capital. En este informe se realizan los pasos de carga y limpieza de datos que permitirán conseguir un DataFrame consolidado que presente los tipos de datos adecuados para la información contenida en sus columnas para una posterior modelación con procesos Gaussianos

Índice de Contenidos

Índice de Figuras

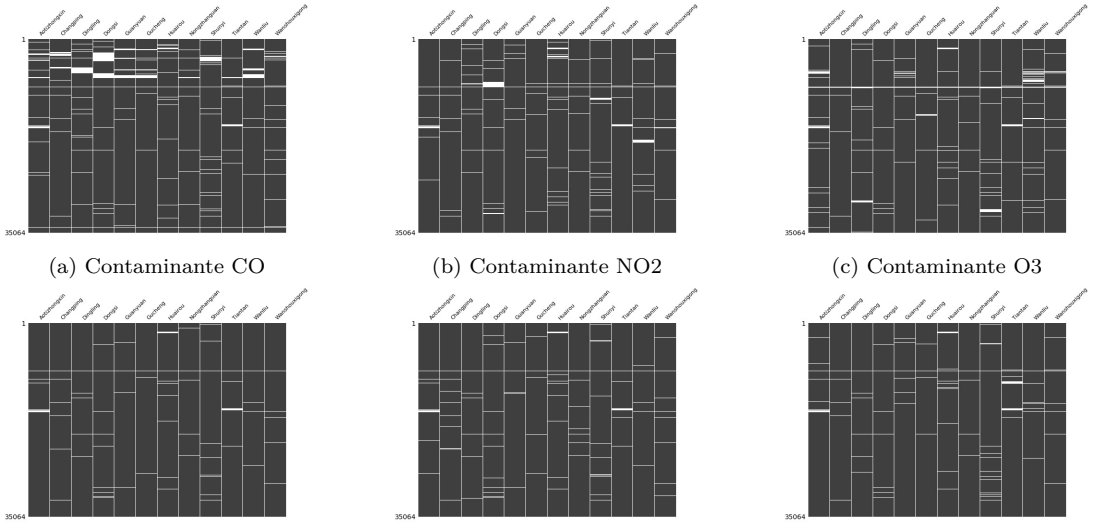
P1) Carga y limpieza de datos

P1.1) Transformación de tiempo a formato Datetime.

Para iniciar se cargan los dataframes asociados a las 12 distintas estaciones de monitoreo, para luego unirlos en un sólo dataframe.
Como cada fila esta esta asociada a una fecha y hora particular, repartidas en distintas columnas, lo siguiente es transformarlas a formato `Datetime` con la función `pandas.to_datetime`, tener los datos temporales en este formato nos permite calcular la cantidad de días y horas de diferencia con una fecha de referencia. Así, usando la resta asociada a este tipo de datos, podemos restar con la fecha 01/03/2013 en formato `Timestamp`, y así obtener la cantidad de dias y horas que de diferencia con respecto a tal fecha.
Antes de pasar al análisis de valores faltantes, se chequeó que, fijando cualquier estación, las fechas estaban ordenadas de manera creciente y que los tiempos en filas consecutivas diferían en una hora para todas las filas.

P1.2) Análisis de valores faltantes.

Para analizar posibles patrones en los datos faltantes, estos se visualizaron para cada contaminante comparando las distintas estaciones. Está primera visualización se realizó con la función `matrix` de la libreria `msno`. Las visualizaciones obtenidas son mostradas en la Figura 1.



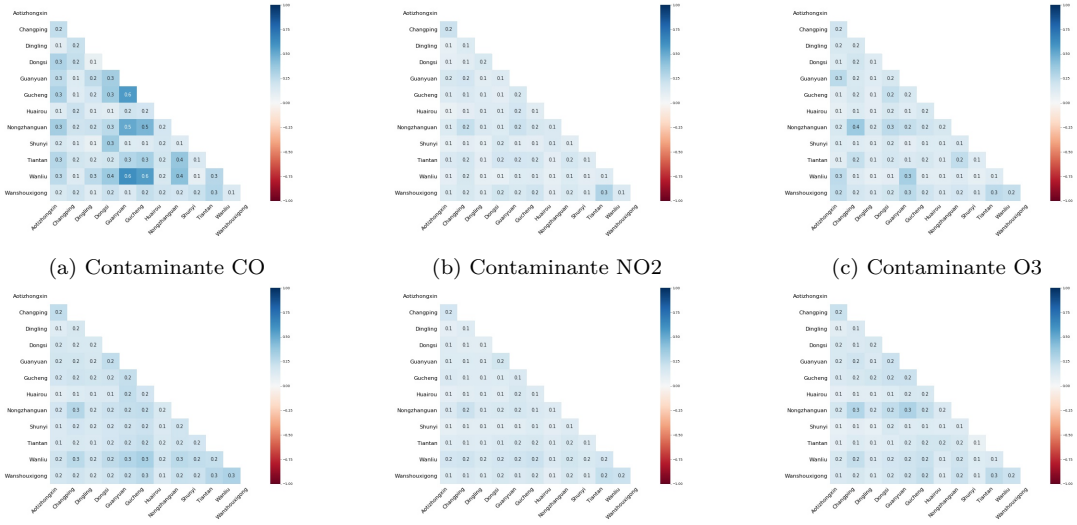
(a) Contaminante CO (b) Contaminante NO2 (c) Contaminante O3

(d) Contaminante PM10 (e) Contaminante PM2.5 (f) Contaminante SO2

Figura 1: Valores faltantes para distintos contaminantes

En estas visualizaciones se puede observar primero, que hay una fecha en que todos los contaminantes y estaciones tienen valores faltantes, y segundo, de manera mas particular, que los valores faltantes asociados a PM10 parecen estar asociados también a los otros contaminantes. Aparte de esto, el resto de valores faltantes parecen estar uniformemente distribuidos entre estaciones, y que en general las fechas en las hay valores faltantes para cada contaminante son distintas en cada estación. Esta segunda observación fue comprobada calculando el porcentaje de valores faltantes de

PM10 que coincidían en los otros contaminantes. Como en todos los contaminantes este porcentaje fue mayor que 77 %, la observación fue acertada.
Debido a que los valores faltantes parecían pocos en cada contaminante, se calculó el porcentaje de estos valores para cada estación, con el fin de no eliminar ningún contaminante. Los porcentajes fueron todos menores que 5 %, siendo el máximo y el mínimo porcentaje alcanzados por los contaminantes CO (4.9 %) y PM10 (1.5 %). Consecuentemente, se continúa el análisis de valores faltantes visualizando las correlaciones entre estaciones sin eliminar ningún contaminante.
Para visualizar las correlaciones, se usa la función `heatmap` de la librería `msno`. Las visualizaciones obtenidas son mostradas en la Figura 2.



(a) Contaminante CO (b) Contaminante NO2 (c) Contaminante O3

(d) Contaminante PM10 (e) Contaminante PM2.5 (f) Contaminante SO2

Figura 2: Correlaciones de valores faltantes para distintos contaminantes

En estas correlaciones se puede observar que en general son bajas y positivas, por lo que, en caso de tener correlaciones altas entre ciertas estaciones para algún contaminante, se podría justificar justificar un llenado de valores faltantes calculando el valor del contaminante en las otras estaciones (que por las correlaciones bajas en los valores faltantes, con alta probabilidad se puede encontrar el valor para ese tiempo en otra estación).

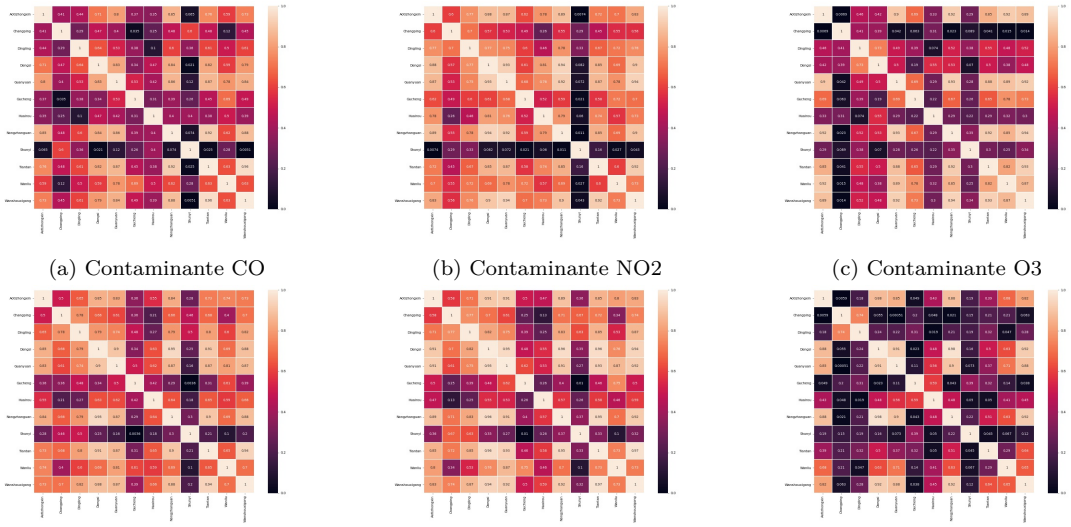
P1.3) Llenado de valores faltantes con interpolación.

En esta sección se llena los valores faltantes en todas las estaciones con el método `Linear`, que puede ser justificado por la tendencia suave que deberían tener las mediciones de un contaminante cualquiera.

P1.4) Correlaciones de contaminantes entre estaciones

Para observar las correlaciones en los distintos contaminantes, se visualizan los heat-maps de las matrices de correlación. Las visualizaciones obtenidas se observan en la figura 3.

Se observa que el contaminante SO2 tiene una gran cantidad de colores oscuros, por lo que se presume que tiene el menor promedio de correlacion, con lo que se procede a calcular los promedios



(a) Contaminante CO (b) Contaminante NO2 (c) Contaminante O3

(d) Contaminante PM10 (e) Contaminante PM2.5 (f) Contaminante SO2

Figura 3: Correlaciones para distintos contaminantes

de correlación. Distinto a la observación, el menor promedio de correlación fue el de NO2 con 0.76. Y para este último contaminante, las 3 estaciones con mayor promedio fueron: Guanyuan (0.79), Wanshouxigong (0.79) y Nongzhanguan (0.78).

P1.5) Conclusión

En este reporte se pudo analizar el análisis de valores faltantes, que permite a posteriori justificar un llenado de datos con interpolación, además de una comprensión mayor del data-set en cuestión. Partes como la descomposición de la señal y la modelación del proceso Gaussiano no fueron puestas en este reporte por no tener suficiente justificación ni análisis.