Taller 2 - Gestión de datos

Calidad del aire - Estación La Flora Santiago de Cali

Ronald Fernando Rodríguez Barbosa
Maestría en Ingeniería de Sistemas y Computación
Maestría en Analítica para la Inteligencia de Negocios
Pontificia Universidad Javeriana

18 de Mayo de 2019

Introducción

La **calidad del aire**, se define como la cantidad general de polución presente en un area y como la pureza promedio atmosférica en relación a las medidas de descarga tomadas de una fuente de polución (Gooch 2007). La contaminación del aire, representa un importante riesgo medioambiental para la salud, bien sea en los países desarrollados o en los países en desarrollo ya que es evidenciado en casos en morbilidad por trastornos cerebrovasculares, cánceres de pulmón y neumopatías crónicas y agudas. Por lo tanto, cuanto más bajos sean los niveles de contaminación del aire, mejor será la salud cardiovascular y respiratoria de la población a largo y a corto plazo (OMS 2016).

De los 23 países de América Latina y el Caribe, 18 tienen sus propias regulaciones vigentes en la actualidad relacionadas con la calidad del aire, que son de acceso público en los sitios web oficiales (Morantes et al. 2016). La trazabilidad para tales regulaciones se establece para los contaminantes de criterio (PM10, PM2.5, SO2, NO2, O-3, CO), utilizando como referencia la secuencia histórica de estándares de la Agencia de Protección Ambiental de los Estados Unidos (USEPA 2019) y los valores de referencia de La Organización Mundial de la Salud (OMS-WHO 2019).

En Colombia, el Sistema de Vigilancia de la Calidad del Aire de Cali (SVCAC 2019), opera bajo la coordinación y administración del Departamento Administrativo de Gestión del Medio Ambiente (DAGMA 2019),Grupo de Calidad del Aire. El Sistema de Vigilancia de Calidad del Aire de Santiago de Cali SVCASC fue acreditado en la norma NTC-ISO/IEC 17025 del año 2005 por el IDEAM a través de la Resolución 1328 del 23 de junio de 2018. El SVCASC actualmente funciona con nueve (9) estaciones: La estación La Flora, ubicada en el barrio La Flora en la zona norte; La estación ubicada en el barrio obrero y la estaciuón Ermita ubicada en el barrio San Pedro ambas en la zona centro; la estación transitoria EDB - Navarro ubicada en el barrio Poblado en la zona oriente; la estación base Aérea, ubicada en el acuaparque de la Caña en zona nororiente; la estación Pance ubicadada en la Zona Rural; la estación Univalle ubicada en el barrio Meléndez en la zona sur; la estación compartir ubicada en el Barrio Compartir en la zona oriente y la estación Cañaveralejo ubicada en la estación SITM del MIO en la zona suroccidente.

El presente trabajo, realizará un análisis de la captura de datos de la estación *La Flora*, la cual es una estación automática que reporta información horaria al centro de control del DAGMA. Esta estación, mide los niveles de Material Particulado Menor a 10 micrómetros (PM10), Dióxido de Azufre (SO2), Dióxido de Nitrógeno (NO2), Monóxido de Carbono (CO), Ozono (O3) y variables meteorológicas como velocidad del viento, dirección del viento, temperatura, humedad, radiación solar y precipitación. El proceso de preparación y análisis de los datos seguirá la siguiente estructura:

1. ***Carga y exploración***: Se incluirá el archivo de datos y se identificarán las características del conjunto de datos
2. ***Limpieza de datos***: A partir de la exploración, se definirán los procedimientos para la limpieza de datos para facilitar el análisis
3. ***Creación de la vista minable***: Se establecerán los conjuntos de datos finales y su correspondiente análisis
4. ***Conclusiones e infografía***: Se compilará el conocimiento adquirido y las cifras de interés adquiridas.

1. Carga y exploración

Tablas de resumen

Para los procedimientos de carga y exploración, se emplearán las herramientas RStudio y RapidMiner. Inicialmente, se realiza la carga de archivos con el fin de resumir las características generales de los datos.

```
datos_base_flora<-read.csv(file="data/dataCAFlora.csv",
                 header = TRUE,sep = ",", dec = ".",fileEncoding = "latin1")
str(datos_base_flora,vec.len=0)
```

```
## 'data.frame':    84629 obs. of  12 variables:
##  $ Fecha...Hora             : Factor w/ 75869 levels "01/01/2011 01:00:00 AM",..: NULL ...
##  $ PM10..ug.m3.             : Factor w/ 1390 levels "0.1","0.2","0.3",..: NULL ...
##  $ SO2...ug.m3.             : Factor w/ 3968 levels "0.03","0.05",..: NULL ...
##  $ NO2...ug.m3.             : Factor w/ 5036 levels "0.02","0.13",..: NULL ...
##  $ CO...ug.m3.              : Factor w/ 8027 levels "1000.10","1000.35",..: NULL ...
##  $ O3...ug.m3.              : Factor w/ 6397 levels "0","0.0","0.02",..: NULL ...
##  $ Vel.Viento...m.s.        : Factor w/ 63 levels "0","0.1","0.2",..: NULL ...
##  $ Dir.Viento..Grados.      : Factor w/ 3602 levels "0","0.1","0.2",..: NULL ...
##  $ Temperatura..CÂ..        : Factor w/ 172 levels "16.2","16.3",..: NULL ...
##  $ Humedad....              : Factor w/ 708 levels "100","100.3",..: NULL ...
##  $ Radiacion.Solar..Watt.M2.: Factor w/ 7176 levels "0","0.1","0.2",..: NULL ...
##  $ Lluvia..mm.              : Factor w/ 182 levels "#","0","0.1",..: NULL ...
```

El conjunto de datos contiene un total 84.629 observaciones con 12 variables. La descripción de las variables se relaciona a continuación:

Variable	Tipo de variable	Descripción
Fecha & Hora	Fecha	Fecha y hora de la captura del senso de polutantes
PM10 (ug/m3)	Contínua	Concentración de Material Particulado Menor a 10 micrómetros
SO2 (ug/m3)	Contínua	Concentración de Dióxido de Azufre
NO2 (ug/m3)	Contínua	Concentración de Dióxido de Nitrógeno
CO (ug/m3)	Contínua	Concentración de Monóxido de Carbono
O3 (ug/m3)	Contínua	Concentración de Ozono
Vel Viento (m/s)	Contínua	Velocidad del viento en metros por segundo
Dir Viento (Grados)	Contínua	Dirección del viento
Temperatura (C°)	Contínua	Temperatura en grados celsius
Humedad (%)	Contínua	Porcentaje de humedad
Radiacion Solar	Contínua	Radiación Solar (Watt/M2)
Lluvia (mm)	Contínua	Cantidad de precipitaciones

Resumen de frecuencia de valores en las variables

Con el fin de identificar los posibles valores en las variables, la frecuencia de dichos valores de forma general y detectar valores ausentes se realiza la siguiente exploración mediante la exposición y resumen. Con el fin de facilitar su visualización se realiza una partición por cada 3 variables, utilizando la función summary del lenguaje R.

En las siguientes listas, se puede apreciar una cantidad significativa de registros con ausencia de valor numérico o valores en cero, con diferentes magnitudes de frecuencia entre las diferentes variables. Según los boletines emitidos por la SVCASC, dichos datos faltantes pueden estar relacionados a las anomalías que se dan en las estaciones de monitoreo, tales como: Fallas en los equipos, falta de energía eléctrica en la zona, hurto de

equipos o cableado, mantenimiento o cambio de equipos y la inclusión o exclusión de algunos contaminantes o variables meteorológicas. Por otra parte, se puede evidenciar que existe más de un registro con misma fecha y hora, lo que puede sugerir un proceso de limpieza en el que se excluirían dichos registros duplicados. Este procedimiento, se verá con más detalle en la sección 2.

```
summary(datos_base_flora[1:3])
```

```
##                  Fecha...Hora    PM10..ug.m3.    SO2...ug.m3.
##  01/01/2011 01:00:00 AM:    2   ND     :17551   ND     :64788
##  01/01/2011 01:00:00 PM:    2   38     :  421   9.56   :   25
##  01/01/2011 02:00:00 AM:    2   48     :  374   11.09  :   24
##  01/01/2011 02:00:00 PM:    2   27     :  344   5.26   :   24
##  01/01/2011 03:00:00 AM:    2   32     :  342   6.77   :   24
##  01/01/2011 03:00:00 PM:    2   39     :  342   8.57   :   24
##  (Other)               :84617   (Other):65255   (Other):19720
```

```
summary(datos_base_flora[4:6])
```

```
##   NO2...ug.m3.    CO...ug.m3.     O3...ug.m3.
##  ND     :58116   ND     :69992   ND     :53828
##  18.48  :   25   1956.56:    9   0.0    :  640
##  20.91  :   25   1100.01:    8   0.5    :  238
##  20.26  :   23   1245.79:    8   0.6    :  207
##  15.03  :   21   1493.70:    8   0.1    :  206
##  22.82  :   21   1526.33:    8   4.3    :  189
##  (Other):26398   (Other):14596   (Other):29321
```

```
summary(datos_base_flora[7:9])
```

```
##  Vel.Viento...m.s. Dir.Viento..Grados. Temperatura..CÂ..
##  ND     :39017     ND     :39017       ND     :39018
##  0.2    : 4434     0      :   52       22.5   :  702
##  0.3    : 4184     33.6   :   46       22.8   :  644
##  0.4    : 3811     48.8   :   41       21.9   :  642
##  0.1    : 3528     272.6  :   40       22.4   :  639
##  0.5    : 3342     87.1   :   39       22.2   :  634
##  (Other):26313     (Other):45394       (Other):42350
```

```
summary(datos_base_flora[10:12])
```

```
##   Humedad....    Radiacion.Solar..Watt.M2.  Lluvia..mm.
##  ND     :39017   ND     :26960             0      :72107
##  81.5   :  139   0      :26733             ND     : 7018
##  81.9   :  138   665    :  821             0.25   : 1874
##  78.9   :  133   664    :  587             0.51   :  563
##  79.9   :  132   0.1    :  414             0.76   :  344
##  78.7   :  131   666    :  379             1.02   :  280
##  (Other):44939   (Other):28735             (Other): 2443
```

Métricas de tendencia central

histogramas

Métricas de dispersión

Diagramas de caja

```r
library(psych)
estadisticas <- describe(datos_base_flora)
estadisticas
```

```
##                             vars     n     mean       sd median  trimmed
## Fecha...Hora*                  1 84629 37914.01 21902.28  37900 37912.40
## PM10..ug.m3.*                  2 84629   836.02   376.87    766   845.50
## SO2...ug.m3.*                  3 84629  3503.71  1044.38   3968  3799.75
## NO2...ug.m3.*                  4 84629  4039.34  1631.06   5036  4341.53
## CO...ug.m3.*                   5 84629  7319.94  1835.35   8027  7874.03
## O3...ug.m3.*                   6 84629  5094.69  2064.35   6397  5516.83
## Vel.Viento...m.s.*             7 84629    33.57    26.82     21    33.92
## Dir.Viento..Grados.*           8 84629  2626.20  1200.02   3346  2794.87
## Temperatura..CÂ..*             9 84629   118.98    53.31    124   123.01
## Humedad....*                  10 84629   547.83   183.98    628   571.94
## Radiacion.Solar..Watt.M2.*    11 84629  3504.55  3140.66   3385  3483.69
## Lluvia..mm.*                  12 84629    19.26    51.23      2     2.55
##                                 mad min   max range  skew kurtosis    se
## Fecha...Hora*               28123.44   1 75869 75868  0.00    -1.20 75.29
## PM10..ug.m3.*                 354.34   1  1390  1389  0.08    -0.83  1.30
## SO2...ug.m3.*                   0.00   1  3968  3967 -2.18     3.28  3.59
## NO2...ug.m3.*                   0.00   1  5036  5035 -1.21    -0.24  5.61
## CO...ug.m3.*                    0.00   1  8027  8026 -2.61     5.57  6.31
## O3...ug.m3.*                    0.00   1  6397  6396 -1.33     0.30  7.10
## Vel.Viento...m.s.*             28.17   1    63    62  0.06    -1.91  0.09
## Dir.Viento..Grados.*          378.06   1  3602  3601 -0.84    -0.82  4.13
## Temperatura..CÂ..*             69.68   1   172   171 -0.31    -1.54  0.18
## Humedad....*                  117.13   1   708   707 -0.75    -0.81  0.63
## Radiacion.Solar..Watt.M2.*   5017.12   1  7176  7175  0.04    -1.78 10.80
## Lluvia..mm.*                    0.00   1   182   181  2.75     5.71  0.18
```

Análisis de las visualizaciones

Análisis de Correlación

2. Limpieza de datos

Reconocimiento y tratamiento de atributos con valores únicos o distintos

Reconocimiento y tratamiento de atributos con valores faltantes

Reconocimiento y tratamiento de atributos con valores atípicos

Reconocimiento y tratamiento de registros atípicos

Reconocimiento y tratamiento de atributos redundantes

3. Creación de la vista minable

Generación de variables derivadas tipo 1 y 2

Normalización de al menos un atributo

Discretización de al menos un atributo

Numerización 1 a n de al menos un atributo

4. Conclusiones e Infografía

Referencias

DAGMA, Departamento Administrativo de Gestión del Medio Ambiente. 2019. "Sitio Oficial - Departamento Administrativo de Gestión Del Medio Ambiente." http://www.cali.gov.co/dagma/.

Gooch, Jan W., ed. 2007. "Ambient Air Quality." In *Encyclopedic Dictionary of Polymers*, 48–48. New York, NY: Springer New York. doi:10.1007/978-0-387-30160-0_522.

Morantes, Giobertti, Narciso Perez, Rafael Santana, and Gladys Rincon. 2016. "A REVIEW OF THE REGULATORY INSTRUMENTS FOR AIR QUALITY AND ATMOSPHERIC MONITORING SYSTEMS: LATIN AMERICA AND THE CARIBBEAN." *INTERCIENCIA* 41 (4): 235–42.

OMS, Organizacion Mundial de la Salud. 2016. "Calidad Del Aire Ambiente Y Salud." http://origin.who.int/mediacentre/factsheets/fs313/es/.

OMS-WHO, Organizacion Mundial de la Salud. 2019. "Sitio Oficial - Organización Mundial de La Salud." https://www.who.int/es/home/.

SVCAC, Sistema de Vigilancia de Calidad del Aire de Cali. 2019. "Sitio Oficial - Sistema de Vigilancia de Calidad Del Aire de Cali." http://www.cali.gov.co/dagma/publicaciones/38365/sistema_de_vigilancia_de_calidad_del_aire_de_cali_svcac/.

USEPA, Agencia de Protección Ambiental de los Estados Unidos. 2019. "Sitio Oficial - Agencia de Protección Ambiental de Los Estados Unidos." https://www.epa.gov/.