PM2.5 Delhi, data sources

Maëlle Salmon

23 janvier 2016

Maybe it's nicer to have a readable document with the analysis!

So, the goal is to compare historic PM2.5 values for Delhi as found on the CPCB website by Eric Dodge to values queried from OpenAQ.

# Load packages

library("readr")  
library("lubridate")  
library("dplyr")  
library("Ropenaq")  
library("ggplot2")

# Check available locations for Delhi on OpenAQ

Ropenaq::locations(city="Delhi", parameter="pm25")

## Source: local data frame [5 x 12]  
##   
## location locationURL city  
## (fctr) (chr) (fctr)  
## 1 Anand Vihar Anand+Vihar Delhi  
## 2 Mandir Marg Mandir+Marg Delhi  
## 3 Punjabi Bagh Punjabi+Bagh Delhi  
## 4 RK Puram RK+Puram Delhi  
## 5 US Diplomatic Post: New Delhi US+Diplomatic+Post%3A+New+Delhi Delhi  
## Variables not shown: cityURL (chr), country (fctr), count (int),  
## sourceName (fctr), firstUpdated (time), lastUpdated (time), parameters  
## (fctr), latitude (dbl), longitude (dbl).

# we'll use only the 4 first ones since the first one  
# is US embassy data  
locationsDelhi <- Ropenaq::locations(city="Delhi",   
 parameter="pm25")[1:4,]

# Load the CPCB historic data

dataCPCB <- readr::read\_csv("cpcb\_ambient\_panel.csv")  
# change this name for compatibility with Open AQ name  
dataCPCB$station[dataCPCB$station=="R K Puram"] <- "RK Puram"  
# filter the locations we have with OpenAQ  
dataCPCB <- dplyr::filter(dataCPCB,  
 station %in% locationsDelhi$location)  
# now off to translating date  
# I am too lazy for finding something more elegant  
dataCPCB$dt\_clean <- gsub("apr", "-04-", dataCPCB$dt\_clean)  
dataCPCB$dt\_clean <- gsub("may", "-05-", dataCPCB$dt\_clean)  
dataCPCB$dt\_clean <- gsub("jun", "-06-", dataCPCB$dt\_clean)  
dataCPCB$dt\_clean <- gsub("jul", "-07-", dataCPCB$dt\_clean)  
dataCPCB$dt\_clean <- gsub("aug", "-08-", dataCPCB$dt\_clean)  
dataCPCB$dt\_clean <- gsub("sep", "-09-", dataCPCB$dt\_clean)  
dataCPCB$dt\_clean <- gsub("oct", "-10-", dataCPCB$dt\_clean)  
dataCPCB$dt\_clean <- gsub("nov", "-11-", dataCPCB$dt\_clean)  
dataCPCB$dt\_clean <- gsub("dec", "-12-", dataCPCB$dt\_clean)  
dataCPCB <- dplyr::mutate(dataCPCB,  
 dateLocal=lubridate::dmy\_hms(dt\_clean))  
# name the column differently  
dataCPCB <- dplyr::mutate(dataCPCB,  
 historicValue=reading\_value)  
# drop useless columns  
dataCPCB <- dplyr::select(dataCPCB,  
 - dt\_clean,  
 - date\_r,  
 - monitor\_read,  
 - reading\_value)

# Get Open AQ data

It is not a rapid query but it does not take months. ;-)

# dataOpenAQ <- NULL  
# for (i in 1:length(locationsDelhi)){  
# firstUpdated <- locationsDelhi[i,]$firstUpdated  
# locationURL <- locationsDelhi[i,]$locationURL  
#   
# seqDays <- seq(from=lubridate::ymd(format(firstUpdated, "%Y-%m-%d")),  
# to=lubridate::ymd("2015-12-31"),  
# by="1 day")  
# seqDays <- format(seqDays, "%Y-%m-%d")  
# for(i in 1:(length(seqDays)-1)){  
# dataOpenAQTemp <- try(Ropenaq::measurements(location=locationURL,  
# parameter="pm25",  
# limit=1000,  
# date\_from=seqDays[i],  
# date\_to=seqDays[i+1]), silent=TRUE)  
# print(seqDays[i])  
#   
# if(class(dataOpenAQTemp)[1]!="try-error"){  
# dataOpenAQ <- rbind(dataOpenAQ,  
# dataOpenAQTemp)  
# }  
#   
# }  
#   
#   
# }  
# # might be useful later  
# dataOpenAQ <- unique(dataOpenAQ)  
# save(dataOpenAQ, file="dataOpenAQ.RData")  
# write.table(dataOpenAQ, row.names=FALSE, file="dataOpenAQ.csv",  
# sep=",")  
load("dataOpenAQ.RData")

Put these data in shape.

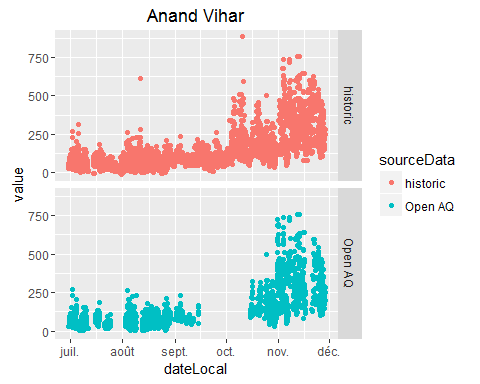
dataOpenAQ <- dplyr::mutate(dataOpenAQ,  
 openAQValue=value,  
 station=location)  
dataOpenAQ <- dplyr::select(dataOpenAQ,  
 dateLocal,  
 station,  
 openAQValue)

# Comparison

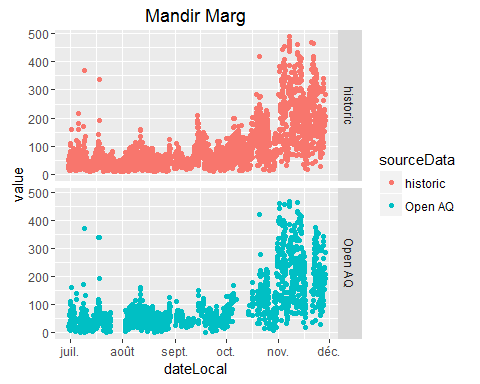
This is the really interesting part I guess.

for (stationNow in levels(as.factor(dataOpenAQ$station))){  
 print(stationNow)  
   
 # filter only data for the station  
 dataTempCPCB <- dataCPCB[dataCPCB$station==stationNow,]  
 dataTempOpenAQ <- dataOpenAQ[dataOpenAQ$station==stationNow,]  
   
 # now filter only dates with data from both sources  
 minDate <- min(dataTempOpenAQ$dateLocal)  
 maxDate <- max(dataCPCB$dateLocal)  
   
 dataTempCPCB <- dplyr::filter(dataTempCPCB,  
 dateLocal>=minDate)  
  
 dataTempOpenAQ <- dplyr::filter(dataTempOpenAQ,  
 dateLocal<=maxDate)  
   
 # now combine both data sets  
 dataTempCPCB <- dplyr::mutate(dataTempCPCB,  
 sourceData="historic",  
 value=historicValue)%>%  
 dplyr::select(dateLocal,  
 value,  
 sourceData)  
 dataTempOpenAQ <- dplyr::mutate(dataTempOpenAQ,  
 sourceData="Open AQ",  
 value=openAQValue)%>%  
 dplyr::select(dateLocal,  
 value,  
 sourceData)  
 dataForPlot <- rbind(dataTempCPCB, dataTempOpenAQ)  
   
 p <- ggplot() +   
 geom\_point(data=dataForPlot,  
 aes(x=dateLocal, y=value, col=sourceData))+   
 ggtitle(stationNow)+ facet\_grid(sourceData ~ .)  
 print(p)  
}

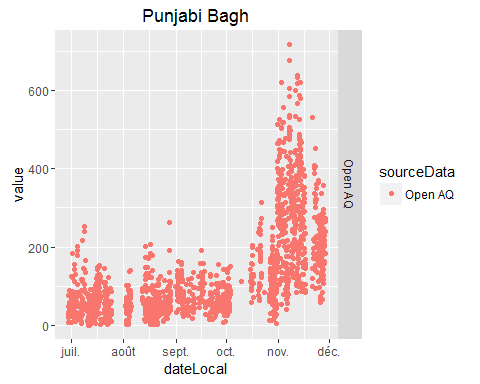
## [1] "Anand Vihar"



## [1] "Mandir Marg"



## [1] "Punjabi Bagh"



## [1] "RK Puram"

