Data Analysis of EPA CO Daily Summary

Daniel Paliura

5/15/2021

Purpose

This document drawn up to research the data in EPA CO Daily Summary data set. This research must set up the ground for mathematical model of forecasting program. Here I have to find which factors affect on CO amounts measured. For example, I expect that location of sites and probably measurements method affect values measured.

I will try to answer questions asked in *Exploratory analysis*, they are bold font. And additionally to find some patterns in data.

The main goal is to determine which data I can use in forecasting and how models for different factors could differs.

Questions

1. How amount of measuring sites differs in different years?
2. Does number of monitors at same sites (unique POC count) changes in time?
3. Why `poc` values 6, 7, and 8 aren't present?
4. Why there are more values with `poc` value 9 than for values 4 and 5?
5. Whether event at some day is written into each observation at same day?
6. Whether data significantly differs by event type factor (regression analysis)?
7. Do events change values in perspective?
8. Would exceptional event presence increases forecasts error compared to forecasting without such event?
9. Why amount of `observation_percent` value equals to 8 is greater than amounts of neighbor values 4, 13, 17, 21, 25?
10. Why feature `arithmetic_mean` contains negative values?
11. Are values of `arithmetic_mean` distributed (log)normally for separate sites/countries/states?
12. Are negative values of `first_max_value` dependent of some factor? Measuring method for example
13. Whether all 1-hour methods has not available AQI?
14. Is there any significant differences in measurements distributions between different methods?
15. How do measured values differs in different states?
16. Does distributions significantly different by factor `cbsa_name`?
17. Are measurements made with NDIR method significantly different by factor `method_code` inside groups NDIR and NDIR PHOTOMETRY? And hence can same methods with different codes be merged?
18. What the result of two-way AnoVa on factor `polutant_standard`?

Analysis

Preparations

I use following R packages:

```
library(dplyr, warn.conflicts = FALSE)
library(lubridate, warn.conflicts = FALSE)
```

And read data. I will use here data set connected with codes, so I will have codes of states and counties and also codes. So I won't be forced to restore relations to determine method or state or county.

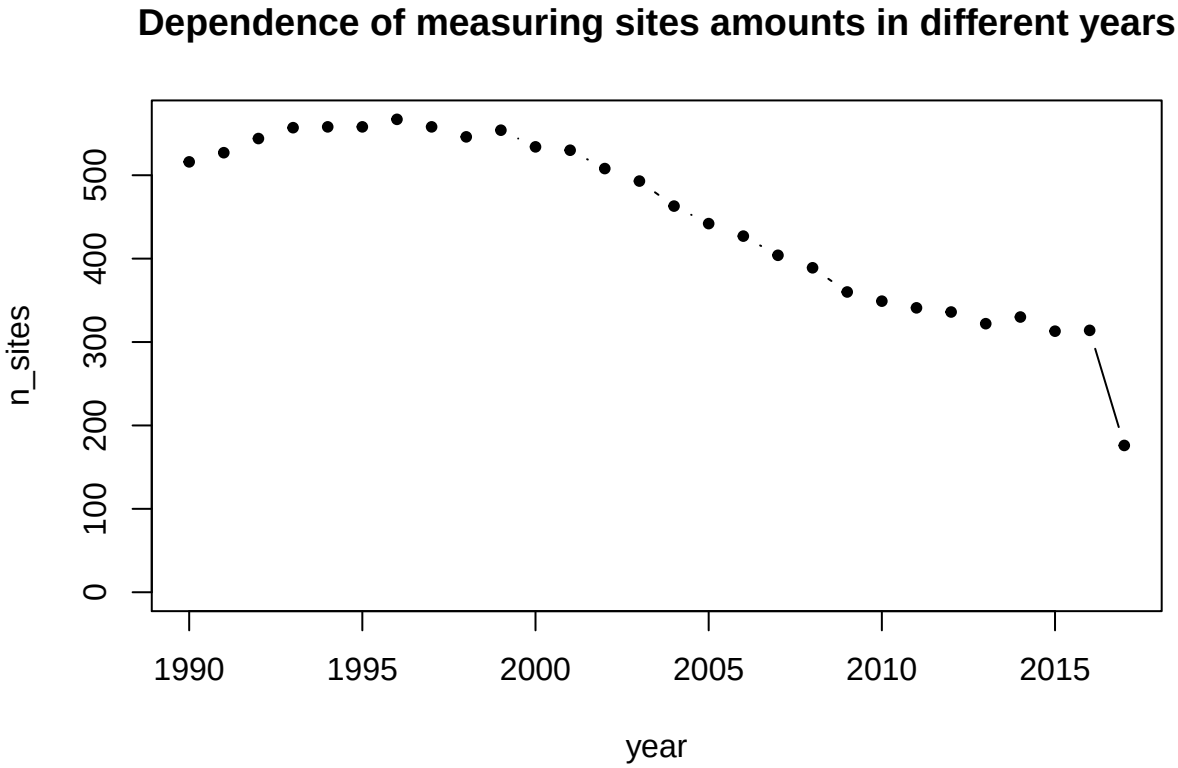
```
folder <- "../data/parted/by_codes/"

na.strs <- c('NA', '', '-')

obs <- read.csv(paste0(folder,"observations.csv"), na.strings=na.strs)[,-1]
sit <- read.csv(paste0(folder,'sites.csv'), na.strings=na.strs)[,-1]
met <- read.csv(paste0(folder,'methods.csv'), na.strings=na.strs)[,-1]

rm(folder, na.strs)
```

Now let's begin answering the questions.

Question 1

How amount of measuring sites differs in different years?

```
## # A tibble: 28 x 2
##     year n_sites
##    <dbl>   <int>
##  1  1990     516
##  2  1991     527
##  3  1992     544
##  4  1993     557
##  5  1994     558
##  6  1995     558
##  7  1996     567
##  8  1997     558
##  9  1998     546
## 10  1999     554
## # ... with 18 more rows
```



Number of measuring sites changing from year to year. First years number of sites was increasing. After 1996 number of sites begun decreasing. It declined by third part after 2010 year.

It would be interesting to see how much sites didn't stop working or worked more than half of whole period.

Question 2

Does number of monitors at same sites (unique POC count) changes in time?

To answer this question I have to group all observations by sites and dates and count unique `poc`.

```
## `summarise()` has grouped output by 'state_code', 'county_code', 'site_num', 'datum'. You can overri

## `summarise()` has grouped output by 'state_code', 'county_code', 'site_num'. You can override using
```

```
## Preview of monitor amounts summary per sites.
## n_min_monitors is minimum number of monitors at the site over the entire period.
## n_max_monitors is maximum number of monitors at the site over the entire period.
```

```
## # A tibble: 6 x 6
## # Groups:   state_code, county_code, site_num [5]
##   state_code county_code site_num datum n_min_monitors n_max_monitors
##        <int>       <int>    <int> <chr>          <int>          <int>
## 1          1          73       23 WGS84              1              1
## 2          1          73       25 NAD27              1              1
## 3          1          73       28 WGS84              1              1
## 4         80           6        7 NAD83              1              1
## 5         80           6        7 WGS84              1              1
## 6         80          26     8012 WGS84              1              1
```

```
## Unique values of n_min_monitors:

## [1] 1

## Unique values of n_max_monitors:

## [1] 1 2 3

## Next table shows how many sites had had which minimum and maximum of monitors

## `summarise()` has grouped output by 'n_min_monitors'. You can override using the `.groups` argument.

## # A tibble: 3 x 3
## # Groups:   n_min_monitors [1]
##   n_min_monitors n_max_monitors sites_amount
##            <int>          <int>        <int>
## 1              1              1         1283
## 2              1              2           14
## 3              1              3            2
```

So now we know that **number of monitors is changeable, but it's a rare phenomenon**. Only 14 of 1299 sites had had maximum 2 monitors through entire period. And only 2 sites had had maximum 3 monitors measuring CO. It's pretty small amount of monitors.

Question 3

Why POC values 6, 7, and 8 aren't present?

```
## Unique POC values for sites with single monitor all the time:

## [1] 1 2 3 4

## `summarise()` has grouped output by 'state_code', 'county_code', 'site_num'. You can override using t

## # A tibble: 16 x 8
## # Groups:   state_code, county_code, site_num [16]
##    state_code county_code site_num datum    poc1  poc2  poc3  poc4
##         <int>       <int>    <int> <chr>   <int> <int> <int> <int>
##  1          5         119        7 WGS84       1     2    NA    NA
##  2          6          19        8 NAD83       1     3    NA    NA
##  3          6          25        5 WGS84       1     3    NA    NA
##  4          6          37     1103 WGS84       1     9    NA    NA
##  5          6          65     8001 WGS84       1     9    NA    NA
##  6          6          77     1002 WGS84       1     3    NA    NA
##  7          6          85        4 NAD83       1     2    NA    NA
##  8          8          41       15 WGS84       1     2    NA    NA
##  9         15           3       10 WGS84       1     2     3    NA
## 10         30          31       17 WGS84       1     5    NA    NA
## 11         37         119       34 NAD27       1     2    NA    NA
## 12         37         119       41 WGS84       1     2     3     4
## 13         50           7        8 UNKNOWN     1     2    NA    NA
## 14         50           7        9 UNKNOWN     1     2     3    NA
## 15         56          39     1012 WGS84       1     2    NA    NA
## 16         80           2        1 NAD27       1     2    NA    NA
```

```
##   state_code        state_name
## 1          6        California
## 2         30           Montana
## 3         37    North Carolina
## 4         80 Country Of Mexico
## 5          8          Colorado
## 6         15            Hawaii
## 7          5          Arkansas
## 8         50           Vermont
## 9         56           Wyoming
```

```
## Methods used at monitors with POC 9:
```

```
##   method_code                                                  method_name
## 1          NA                                                         <NA>
## 2         593 INSTRUMENTAL - Gas Filter Correlation Teledyne API 300 EU
##   pollutant_standard
## 1     CO 8-hour 1971
## 2     CO 1-hour 1971
```

Numbers from 1 through 4 present for sites with single monitor all time and are common for sites with many monitors for CO measuring.

Also, it was 4 monitors through all time at site number 41 present in table. And 3 sites has 3 unique POC values, while it was just 2 sites with maximum 3 monitors at the same time. I guess, measurements was just transported from one monitor to other.

I guess, sites have about fixed numbers of monitors with corresponding fixed numbers. For example, I guess, there are 9 or more monitors in two sites at California. California is advanced state, they can let such many monitors to measure different values. And monitors with number 9 could both be chosen to measure CO, probably due to good location or any other reasons. One monitor with POC 9 was measuring with uncommon method 'INSTRUMENTAL - Gas Filter Correlation Teledyne API 300 EU'. Probably such a monitor was chosen to experimentalize with method.

I guess presence of second site with CO monitor number 9 is randomness. And **POC numbers 6, 7, and 8 aren't present because it wasn't a case**.

Question 4

Why there are more values with poc value 9 than for values 4 and 5?

POC 9 is more frequent than 5 because POC 5 is pretty randomly appeared and at single site.

POC 9 also was used in California to measure CO with as one of main monitors. Monitor number 4 could be secondary at sites where it present.

Question 5

Whether event at some day is written into each observation at same day?