How do Pollutant Gases Influence Nitric Oxides?

Brandon Cohen

Pollution has become an increasing concern worldwide year in and year out with clear detrimental effects to the atmosphere, water, soil, climate, ecosystem, etc. Understanding the contributions that lead to pollution is crucial for effective pollution control and mitigation strategies. One avenue of concern can be observed from pollutant gases such as nitrogen oxides (NO_x), sulfate (SO_4), and ammonia (NH_4). Nitrogen oxides and SO_4 are predominantly produced from the combustion of fossil fuels, and ammonia comes naturally from the soil but also from agricultural products. Nitrogen oxides are one of the most concerning pollutants as they can react with another pollutant called volatile organic compounds (VOCs) in the atmosphere to form ground-level ozone. Ground-level ozone (O_3) is lower in the atmosphere and is very harmful to breathe. SO_4 and NH_4 are just as bad due to their damaging impact on the environment and ecosystem with acidification of water/soil and contributions to the formation of other dangerous pollutants. In this study, the goal was to focus on the relationship between pollution gases by forming a multiple linear regression model to describe the concentration of TNO_3 (total nitrate) with respect to the concentration of NH_4, NO_3, HNO_3, and SO_4 levels using a dataset from the Environmental Protection Agency. Based on recorded nitrogen oxide production over a set time period, a null hypothesis was formed: NH_4, NO_3, HNO_3, and SO_4 do not influence TNO_3 levels. The alternative hypothesis was at least one of the pollutant gases does significantly influence TNO_3 levels with $\alpha = 0.05$.

Before creating the model, it was important to see the trend of the production of TNO_3 in the atmosphere. Based on other reports of gas production, it was expected that the TNO_3 concentration would be Poisson distributed. Figure 1 shows certain characteristics of a Poisson process with a mean of 1.057 $\mu g/m^3$ with a certain level of right skewness and density of values concentrated between 0 and 2.66 $\mu g/m^3$ which is considerably high and toxic.

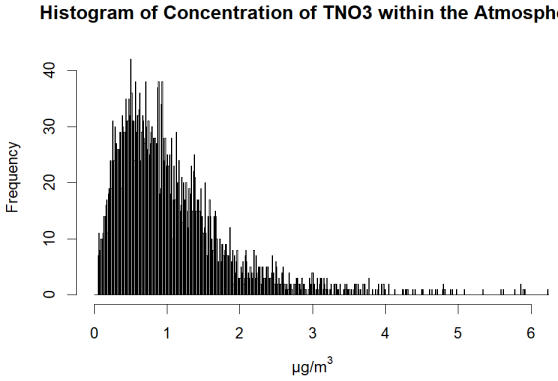

Figure 1. Histogram of Recorded TNO3 in the Atmosphere

In order to build the model, a key assumption was made which was that nitrogen-containing pollutants recorded in the study will influence TNO_3 levels. Adding explanatory variables was done using analysis of single models, but after investigating each model, a selection of some variables was added to the model to better fit the data, which was repeated until the best model was formed. In order to fix the model, parameter testing was done

by evaluating the p-value to 0.05, but after doing this and residual analysis, there were clearly some parameters missing. This is when SO_4 was tested within the model, but it was found that squaring SO_4 provided better results due to the residual parabolic behavior. The model was almost done, but there were still interaction terms in play that were not considered but are well known in chemsitry. The most obvious interaction was between SO_4 and HNO_3 because they are known to interact within the atmosphere forming secondary pollutants. Other interactions of pollutants were tested, but they would consistently decrease the F-statistic and/or the R^2_{adj}. The final model was a second-order quantitative model: $TNO_3 = 0.0000131SO_4^2 + 0.0000424SO_4*HNO_3 + 0.984HNO_3 + NO_3 - 5.72NH_4$. When this model was formed, the first test performed was the F-statistic which was phenomenal with a value of $2.07*10^9$ with a p-value of $<2.2*10^{-16}$. This implied that there was some merit to this model because at least one of the parameters was not equal to zero, but the R^2_{adj} was even better with a value of 1 meaning that the model fits the model incredibly well even with taking into account the number of parameters of the model. While it was almost too perfect to be true to explain the levels of TNO_3 as a function of the NH_4, NO_3, HNO_3, and SO_4 levels, residual analysis in the form of a Quantile-Quantile plot was performed showing there was light-tailedness with more values at both extremes, but overall, they were very good residuals. The validity of the model was further confirmed with the calculation of the residual standard error of an incredibly low value of 0.000502.

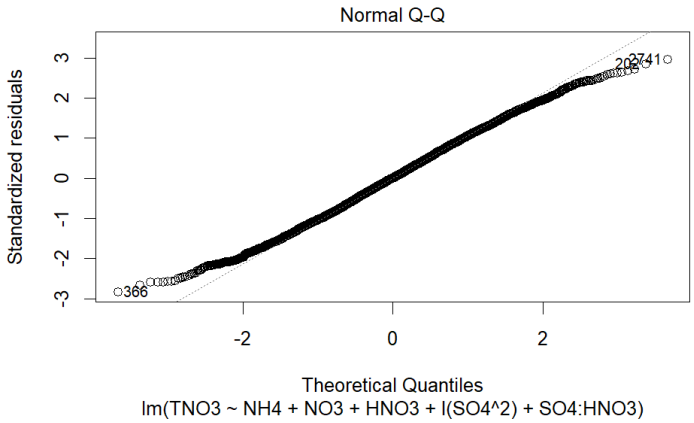

Figure 2. Normal Q-Q Test of Residuals

Overall, the variation of TNO_3 can be explained with 100% of the model showing a very clear connection that the levels of one pollutant are dependent on the levels of other pollutant gases. The only solution to this problem is to take a global approach to minimize the levels of all pollutants because they will continue to be cancerous for our environment. In the future, it would be important to analyze potential exponential behaviors of pollutant levels explained by other pollutant gases. There was no exponential behavior that could be found even during residual analysis, but this does not mean there is no exponential behavior. The importance of finding exponential behavior of pollutant gases is very important because it provides more conclusive proof as to why the world needs to take a global approach to get rid of pollutants as exponential behavior means the atmosphere will continue to break down at a very concerning high rate leading to further destruction of the geosphere and ecosystems around the world.

Appendix

The results of this analysis provide valuable insight into the relationship between pollution gases with respect to TNO3 levels which contributes to the understanding of air pollution dynamics and interactions giving rise to better development of targeted pollution control strategies.

Sample Size = 4092

The data was gathered by the United States Environmental Protection Agency

Variables:

- gas (data set)
- Gases and Minerals
    - HNO3
    - Ca
    - K
    - Mg
    - Na
    - NH4
    - NO3
    - SO2
    - SO4
    - TNO3
- df (data frame contained all the gases and minerals)
- Linear Models
    - lmod1
    - lmod2
    - lmod3
    - lmod4
    - lmod (Final Model)

R Code:

```
# https://catalog.data.gov/dataset/clean-air-status-and-trends-network-castnet-ozone
# https://www.epa.gov/castnet/castnet-ozone-monitoring
# https://www.epa.gov/castnet/download-data

# Read in the dataset into the variable called gas
gas <- read.csv("C:/Users/bwc07/Downloads/table_export.csv")

# Create function to filter data set
filter_vec <- function(vector) {
  # Create new vector that takes in numbers
  new_vector <- numeric()
```

```
  # Iterate over the whole column
  for (x in vector) {
    # If item in column is empty, skip
    if (x == '-') {
      next
      # If there is a value, continue
    } else {
      # The value at x, convert it into a double
      temp <- as.double(x)
      # Add to the vector called new_vector
      new_vector <- append(new_vector, temp)
    }
  }
  return(new_vector)
}

# Filtered Calcium data, but there was no
# connection found
Ca <- filter_vec(gas$Ca)

# Filtered Nitric acid data
HNO3 = filter_vec(gas$HNO3)

# Filtered Potassium Data, but no connections
# could be found
K = filter_vec(gas$K)

# Filtered Magnesium Data, but no connections could be
# found
Mg = filter_vec(gas$Mg)

#Filtered Sodium Data
Na = filter_vec(gas$Na)

# Filtered Ammonia Data
NH4 = filter_vec(gas$NH4)

# Filtered Nitrate
NO3 = filter_vec(gas$NO3)
```

```
# Filtered Sulfur Dioxide
SO2 = filter_vec(gas$SO2)

# Filtered Sulfate
SO4 = filter_vec(gas$SO4)

# Filtered Total Nitrate
TNO3 = filter_vec(gas$TNO3)

#I am have to further filter data sets so that
# I can make a data frame of 4062 items
Ca <- Ca[-c(4064, 4063)]
HNO3 <- HNO3[-c(4065, 4064, 4063)]
Mg <- Mg[-c(4065, 4064, 4063)]
NH4 <- NH4[-c(4065, 4064, 4063)]
NO3 <- NO3[-c(4065, 4064, 4063)]
SO2 <- SO2[-c(4063)]
SO4 <- SO4[-c(4065, 4064, 4063)]
TNO3 <- TNO3[-c(4065, 4064, 4063)]
df <- data.frame(Ca, HNO3, K, Mg, NH4, NO3, SO2, SO4, TNO3)


# Plotted all histograms of each Gas to find trends of
# their levels. This helped me pick TNO3, because it was
# well described and there were correlations to other gases
# like HNO3 or NH4 that are well known in chemistry.

# Some Notes on some of the histograms:
# Calcium is very right skewed
# Potassium and Na did not deviate much and has a very
# low average concentration in the atmosphere which is expected
# because normally, it is in the ground as a mineral.
# SO2 looks the most like a Gaussian distribution, but it
# is still right skewed.
# Overall, every histogram is right skewed with varying degress
# of how much.
hist(Ca, xlim = c(0, 6), breaks = seq(0, 6, 0.1))
hist(HNO3, xlim = c(0, 3), breaks = seq(0, 3, 0.1))
hist(K, xlim = c(0, 6.3), breaks = seq(0, 6.3, 0.01))
hist(Mg, xlim = c(0, 0.9), breaks = seq(0, 1, 0.01))
```

```
hist(Na, xlim = c(0, 6.6), breaks = seq(0, 6.6, 0.01))
hist(NH4, xlim = c(0, 7), breaks = seq(0, 7, 0.01))
hist(SO2, xlim = c(0, 4), breaks = seq(0, 4, 0.01))
hist(SO4, xlim = c(0, 5), breaks = seq(0, 5, 0.01))
hist(TNO3, main = "Histogram of Concentration of TNO3 within the Atmosphere",
    xlab = expression(plain("µg/m"^"3")), xlim = c(0, 6),
    breaks = seq(0, 7.3, 0.01))

# Where does 95% of the data fall for TNO3?
# Find the range so we know where the min and max are
range(TNO3)
# Find mean of the concentrations
mean(TNO3)
# Find the standard deviation of TNO3 conc.
std_TNO3 <-sd(TNO3)
# Print mean(TNO3) +/- 2*s
print(mean(TNO3) + 2*std_TNO3)
print(mean(TNO3) - 2*std_TNO3)

# I want to find how many values are outside of
# mean(TNO3) +/- 2*s which is only 177 out of 4062
# which is good because most of the data is still
# within range
out_of_range <- sum(TNO3 < 0 | TNO3 > 2.6612)
out_of_range

# Beginning of Model Building for air pollutant TNO3
#TNO3 (Total reactive Nitrogen Oxides includes various
# nitrogen oxides such as NO, NO2, and NOx.

# Nitric acid is a component of acid rain, which can have
# negative impacts on plants, animals, and ecosystems.
# Acid rain occurs when sulfate (SO4) and nitrogen
# oxides (NOx), which are produced by fossil fuel combustion
# like in transportation, and react with atmospheric
# moisture to form sulfuric acid and nitric acid.

# I know any NOx should in theory influence
# TNO3 levels since all NOx are produced back
# and forth in nature.
```

```
# This model had a bad R^2, but the F-statistic was
# very good so I kept this variable in mind when
# building the model but I did not use it immediately
lmod1 <- lm(TNO3 ~ HNO3, data=df)
summary(lmod1)

# Similar situation with NO3 where it has a bad
# R^2 but a very high F-statistic so I also kept
# this variable in mind just in case I could use it
lmod2 <- lm(TNO3 ~ NO3, data=df)
summary(lmod2)

# I have a significantly higher R^2 = 0.70 and an incredibly high
# F-statistic for this model so I know for sure that I
# have a great variable to use in the model
lmod3 <- lm(TNO3 ~ NH4, data=df)
summary(lmod3)

# I also chose to model SO4 because
# to see how it would interact with the data since
# I know SO4 interacts with other pollutants to form
# NOx. It had a very high F-statistic but a low R^2
lmod4 <- lm(TNO3 ~ SO4, data=df)
summary(lmod4)

# After modifying other variables, I find the best fit with the best
# residual analysis and goodness of fit, I get the following:
lmod <- lm(TNO3 ~ NH4 + NO3 + HNO3 + I(SO4^2) + SO4:HNO3, data=df)  ### AMAZING
MODEL
# The F-statistic is insanely high which gives me confidence that at
# least one of the parameters are significant. I then check the R^2
# and find that it fits the data incredibly well. Then, I perform
# two parameter test on NH4 since it has a very high corrlation by itself
# and HNO3:SO4 since it is a known reaction to form NOx, but I was very surprised
# to find such high Pr(>|t|). The most surprising part was when I removed the
# variables, I did not get the same best model, so when I added each one back at a
# time, I gathered the best model.
summary(lmod)

# This shows a great correlation that shows the error follows a normal
```

```
# distribution with slight light tailing at both ends, but this is expected
# especially for any real data set. It will almost never be perfect
plot(lmod, which = 2)

# I also looked at this residual plot, which definitely showed that the
# residuals were leaning more into the the range of 0 to 2, which of course
# should not be the case, but I actually enlarged the plot and looked into the
# residuals that were past the x-axis =3 and I found most of those values were
# outliers so if you remove those values, you get a perfectly random plot of residuals.
plot(lmod, which=1)



# Use of Pearson Coefficient
# Suggests a strong linear relationship between the levels of
# TNO3 and NH4 with 95% confidence
# Since r=0.83, it can be implied that as TNO3 levels increase,
# so does the levels of NH4
cor.test(TNO3, NH4, method='pearson')

# I wanted to find some relationships between the variables
# There are are very clear linear behaviors of NH4 vs. TNO3, NO3 vs. TNO3, NO3 vs. NH4
options(repr.plot.width = 10, repr.plot.height = 10, repr.plot.res = 300)
pairs(df[,c("TNO3", "HNO3", "NH4", "NO3", "SO4")],
      pch=20, # point shape
      lower.panel = NULL # don't show below diagonal
)

# Minimum sum of the squared errors.
SSE = sum((lmod$residuals)^2)     # = 0.001
# Get the degrees of freedom
deg_free = lmod$df
# Root Mean Square aka Residudal Standard Error
s = sqrt(SSE/deg_free)            # = 0.0005

# We expect that the model will roughly predict future
# levels of TNO3 based on the independent gases to
# about ±2s = 0.001 ug/m^3 which is a reasonably small concentration
#±2s = ±2(133.5) = ±267 dollars.will provide a rough approximation to the accuracy with which
the model will predict future values of y for given values of x. Thus, in Example 4.1, we expect
the model to provide predictions of auction price to within about ±2s = ±2(133.5) = ±267 dollars.
```

2*s

```
# This shows a clear relationship that as NH4 levels increase, so does TNO3
# What can also be observed is the fact that if the concentration of NH4 is
# within 0-0.5 ug/m^3, the rate at which TNO3 will not be nearly as great
# as when the concentration of NH4 > 1.0 ug/m^3
plot(TNO3 ~ NH4, data=df, main = 'Realtionship between TNO3 and Ammonia')
abline(lm(TNO3~NH4,data=df), col='red', lwd=2)



# TNO3 = 0.0000131SO4^2 + 0.0000424SO4*HNO3 + 0.984HNO3 + NO3 - 5.72NH4
newdata  = data.frame('SO4'=mean(SO4), 'HNO3'=mean(HNO3), 'NO3'=mean(NO3),
'NH4'=mean(NH4))
#forming a confidence interval with alpha=0.05
predict(lmod, newdata = newdata, interval='confidence', level=0.95)

# forming a prediction interval
predict(lmod, newdata = newdata, interval='prediction', level=0.95)
```