Spring 2024

Analyzing trends in ground-level ozone

Keily Hart
St. Mary's University, khart@mail.stmarytx.edu

Analyzing Trends in Ground-Level Ozone

by

Keily Hart

HONORS THESIS

Presented in Partial Fulfillment of the Requirements for
Graduation from the Honors Program of
St. Mary's University
San Antonio, Texas

Dec 2023

Approved by

Kaitlin Hill

Dr. Kaitlin Hill
Department of Mathematics

Lori Boies

Dr. Lori Boies
Honors Program

Analyzing Trends in Ground-Level Ozone

Keiily Hart

Abstract

In the last few decades, concerns regarding air pollution have led to many new laws and regulations being put into place to mitigate the effects of pollution on the environment and humanity as a whole. This paper analyzes several decades' worth of ground-level ozone readings in six of the largest metropolitan areas in Texas, using data from the United States Environmental Protection Agency (EPA). These regions include the Austin-Round Rock area, Corpus Christi, the Dallas-Fort Worth-Arlington area, El Paso, the Houston-The Woodlands-Sugar Land area, and the San Antonio-New Braunfels area. We identify trends in these readings using the Jonckheere-Terpstra test.

Keywords: Extreme Value Theory; Air Pollution; Clustering; Jonckheere-Terpstra test; Ground Level Ozone; Particulate Matter

Contents

1 Introduction

Air pollution and its effects on the environment, animals, and humanity have been an ever-increasing concern in the past several decades. As humans learn more about our impact on the environment, individuals and institutions strive to reduce that impact. In spite of many efforts to reduce emissions, air pollution remains a significant concern all around the world, not just in developing countries or urban centers. The World Health Organization (WHO) estimates that 99% of people around the world breathe air that is extremely polluted, far exceeding the WHO guidelines [1].

Ground-level ozone is a harmful air pollutant that can cause respiratory problems. This ozone differs from the ozone in the upper atmosphere, which naturally forms a protective layer that shields the Earth from the sun's harmful ultraviolet rays. While upper-atmosphere ozone is beneficial, ground-level ozone harms human health and the environment, and exposure to high levels of ground-level ozone can cause issues with the respiratory system. Even less significant levels of ground-level ozone can cause lasting effects on a person's respiratory system [2]. These issues can include shortness of breath and coughing, as ground-level ozone can make it difficult to breathe. Ground-level ozone is produced from chemical reactions between nitrogen oxides (NOx) and volatile organic compounds (VOCs) in sunlight, such as carbon monoxide (CO) or methane [3]. As a result, ground-level ozone's effects are most noticeable during the summer.

In an effort to determine if the actions of individuals and institutions to reduce air pollution are in vain, as the above WHO statistic might imply, we analyzed ground-level ozone pollution in six major metropolitan areas in Texas. This is not the first study of this kind; in 1989, Richard Smith did a similar analysis of ground-level ozone pollution in Houston, Texas. His research focused on several methods to analyze extreme values [4]. To my understanding, this is the only other such study addressing trends in extreme values of ground-level ozone. The goal of this analysis was to identify decreasing trends in ground-level ozone levels in the six metropolitan areas in Texas over the last forty-three years. In order to determine trends, methods including clustering and the Jonckheere-Terpstra hypothesis test for trend were applied to the data. The Jonckheere-Terpstra hypothesis test was chosen because it is an incredibly robust test whose results are not heavily swayed by outliers. This resistance to outliers is important to consider because the goal of this analysis is to determine simply the presence of a decreasing trend in ground-level ozone readings, not considering the magnitude of such reductions. When compared to other tests, such as the Kendall test and the Mann-Whitney test, the Jonckheere-Terpstra hypothesis test fits the desired analysis the best. Conclusions regarding the presence of trend were drawn at several thresholds.

1.1 The Data

The data for all six locations was gathered by the United States Environmental Protection Agency (EPA). The EPA provides data for core-based statistical areas (CBSA or metropolitan areas). The six areas that were researched were: the Austin- Round Rock CBSA, the Corpus Christi CBSA, the Dallas- Fort Worth- Arlington CBSA, the El Paso CBSA, the Houston- The Woodlands- Sugar Land CBSA, and the San Antonio- New Braunfels CBSA. The data for each location consists of daily ground level ozone readings from the last forty-three years. These locations were chosen because of their size and spread across Texas. All six metropolitan areas are larger than 250 square miles, with the greater Houston area being the largest at just under 10,000 square miles [5]. The smallest area studied was the El Paso area, which is only 259 square miles according to the city of El Paso [6]. These areas all have very large populations, with the greater Houston area again having the largest population at 7.2 million people as of 2020 [5]. According to the US Census Bureau, Corpus Christi has the smallest population, with only 316,239 people [7].

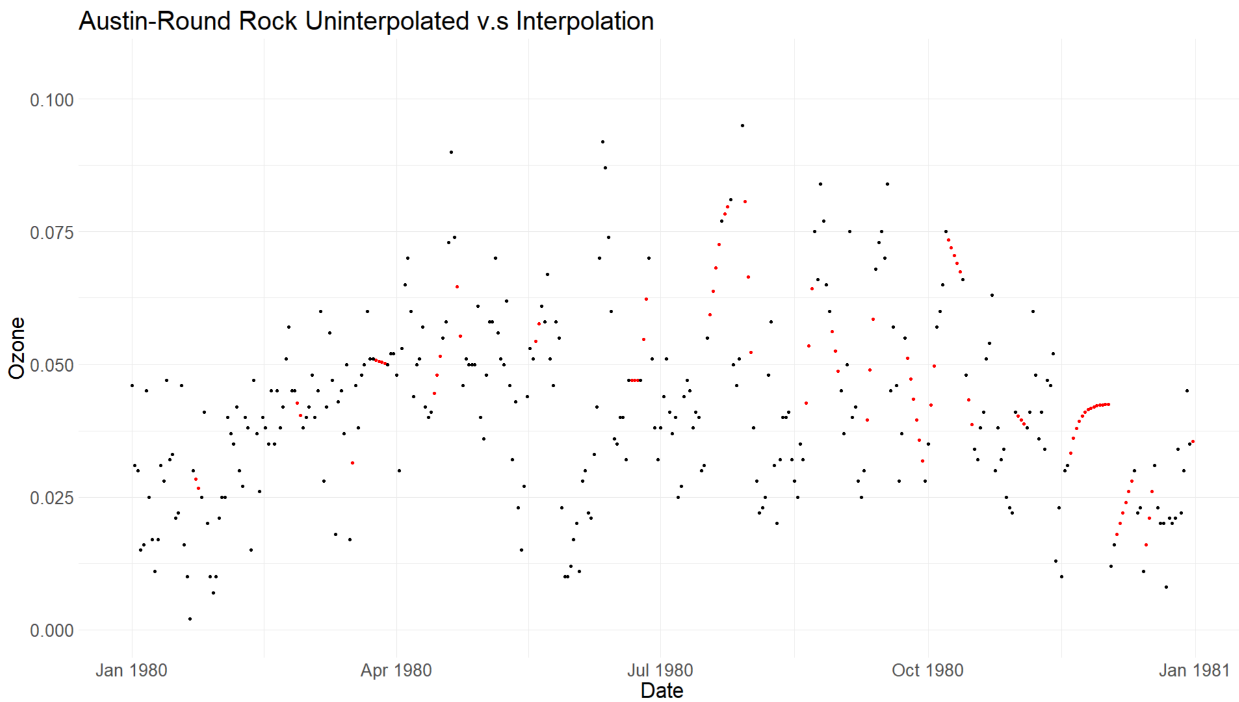

Figure 1: Note that the red dots represent interpolated data, while the black dots represent the raw data.

Most real-world data will unfortunately be missing some data points. In order to accommodate for the missing ozone readings in the data, two methods were used in conjunction with one another. For strings of missing ozone readings less than 10 days long, linear interpolation was used to approximate the missing readings. For strings of missing ozone readings of at least 10 days straight, an Seasonal Autoregressive Integrated Moving Average (SARIMA) model was used to forecast the missing data points, using the data from prior to the first missing reading in that section. For an example of the raw data overlaid with the interpolated data, see Figure 1. This scatter plot allows one to see where the interpolated data fills in the missing data points.

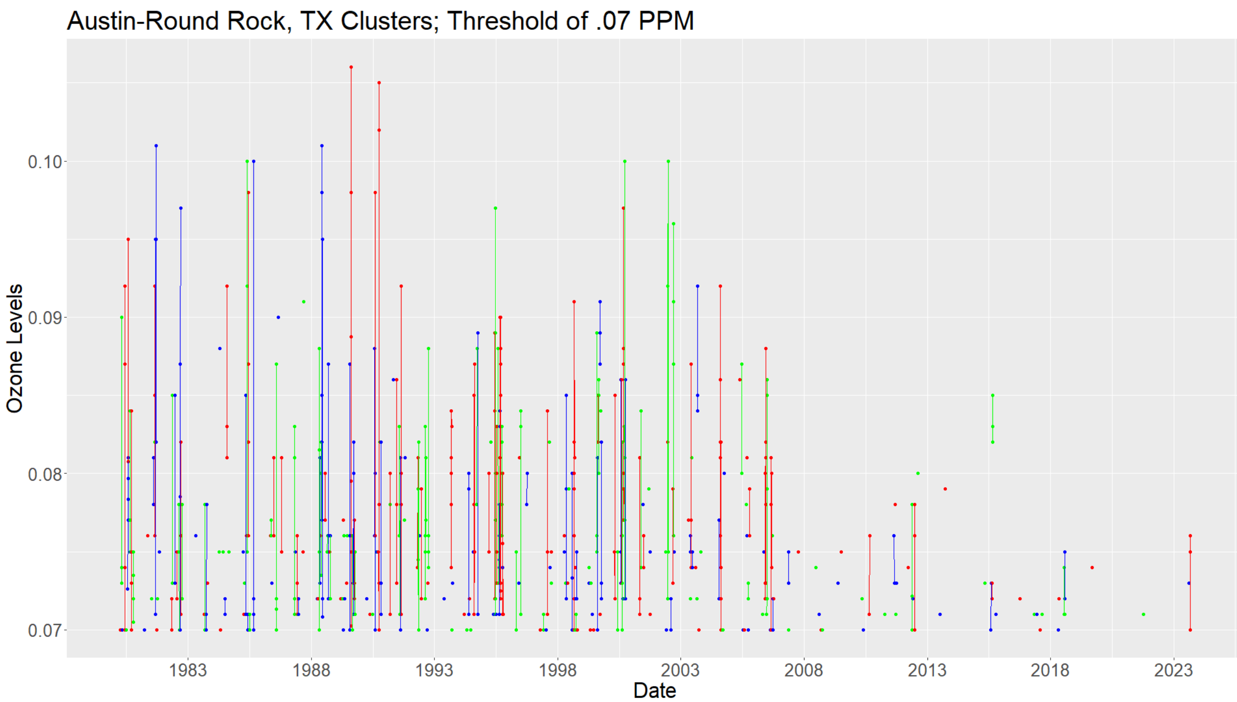

Figure 2

1.2 Clustering

When attempting to effectively identify trends in the data, it is not effective, nor is it efficient, to analyze every data point. One method that is commonly used to analyze trends in ground-level ozone readings is clustering the extreme values of the data set together [4]. This method is seen specifically in the research conducted by Richard Smith in 1989. The first step when clustering is to establish a threshold. Four thresholds were chosen for this analysis, 0.08, 0.075, 0.07, and 0.065. These values were chosen because the United States Environmental Protection Agency has reduced the acceptable level of ground-level ozone four times over the last 43 years. The initial level was 0.12 ppm, but in 1997 that standard was reduced to 0.08 ppm. In 2008, the standard was again reduced to 0.075 ppm. Finally "on October 1, 2015, EPA strengthened the ground-level ozone standard to 0.070 ppm, averaged over an 8-hour period"[8]. The lowest threshold of 0.065 was chosen in an effort to see how increasing the volume of readings provided would affect the results. The threshold of 0.12 was not used in the present analysis, as there was not sufficient data at or above that threshold for the analysis to be performed. The method of clustering provides important insight into the behavior of ground-level ozone pollution in the areas of interest and allows one to track changes in the level of exceedances over time.

After a threshold is established, the clusters can be formed. First, we must establish a cluster interval. For this paper, a cluster interval of 72 hours was chosen [4]. A program was built to analyze the data and identify values over the established threshold, see Appendix B. The program then further analyzes the time between exceedances, establishing a new cluster every time the exceedances are further apart than the cluster interval. If the exceedances are closer together than the cluster interval, they are deemed to be a part of the same cluster. Richard L. Smith's analysis of trend in the extreme value of ground-level ozone in Houston, Texas used this method of clustering [4]. His analysis used the maxima of the clusters while ours will use all of the data points contained in the cluster [4]. This is because the Jonckheere-Terpstra hypothesis test makes use of all the data points in the cluster. Once the clusters are identified, they are assumed (under the null hypothesis) to be independent from one another and to originate from the same distribution. Thus the clusters are effectively independent and identically distributed. For an example of how the clusters appear in the data, see Figure 2. This scatter plot represents the clusters formed in the Austin-Round Rock data at a threshold of 0.07. The clusters rotate is color, beginning with red, then green, then blue.

1.3 Hypothesis Testing

The Jonckheere-Terpstra Hypothesis test for trend has two hypotheses: a null and an alternative. The null hypothesis, H_0, states that no trend is identified in the data analyzed. The alternative hypothesis, H_a, varies depending on what the researcher needs to test for. The alternative hypothesis could test for an increasing trend or a decreasing trend. For the purposes of this analysis, the alternative hypothesis will represent a decreasing trend. The results of this test is a test statistic called JT, and a corresponding $p-value$. Since the

JT test statistic is difficult to interpret, one typically uses the corresponding $p - value$ for statistical analysis. It is up to the researcher to choose a significance level (α) to determine whether or not the null hypothesis will be rejected. The significance level represents the researchers confidence in their rejection or acceptance of the null. The smaller the significance level, the more confident one can be in their rejection of the null hypothesis. The significance level will be compared to the $p - value$ obtained by the test. If the alpha value is smaller than or equal to the $p - value$, the researcher will fail to reject the null hypothesis. If the alpha value is larger than the $p - value$, the researcher will reject the null hypothesis in favor of the alternative hypothesis. Commonly, $\alpha = 0.05$ is used, however, in this analysis a more conservative $\alpha = 0.01$ is used.

2 Development of the Jonckheere-Terpstra Test

2.1 M. G. Kendalls Impact on Tests for Trends

In 1938, M. G. Kendall published his paper on measures of rank correlation. He analyzes methods of establishing r, a representation of this correlation. He begins with an example:

Let

$$A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Establish an "arbitrary ranking"

$$A_1 = \{4, 7, 2, 10, 3, 6, 8, 1, 5, 9\}.$$

Kendall then considers the pairs of values in A such that 4 is paired with each number that may come after it [9]. Thus the ordered pair $(4, 7)$ is valid, as is $(4, 2)$. Kendall then states that for all of these pairs of $(4, x)$ (of which there are nine), a score of 1 is awarded if $x > 4$, and a score of -1 is awarded if $x < 4$ [9]. Thus for the ordered pair $(4, 7)$, a score of positive 1 is awarded, and for the ordered pair $(4, 2)$ a score of -1 is awarded. We do this for

$(4, 10)$ through $(4, 9)$. The nine scores are then added together, totaling +3.

We then repeat this process for $(7, x)$ for all eight numbers that follow 7. This excludes 4 from being x, as 4 comes prior to 7. These eight total to -2. We do this for all nine numbers contained in the set and end up with nine scores,

$$+3,\ -2,\ +5,\ -6,\ +3,\ 0,\ -1,\ +2,\ +1,$$

which sum to +5.

The maximum possible score for this example is 45. This score is given if the numbers are ordered from smallest to biggest, as in A. Kendall then defines a variable r which represents a rank correlation coefficient [9].

$$r = \frac{\text{actual score}}{\text{maximum possible score}} = \frac{5}{45} = .11.$$

Kendall then defines a formula for r that is recursive, using n to represent the number of individuals and Σ is the actual score, such that

$$r = \frac{2\Sigma}{n(n-1)}[9].$$

The larger the number when compared to the maximum possible score for the set, the more 'ordered' the set is. This score will be the foundation for Henry Mann and D. R. Whitney's test, which is the basis of the Jonckheere-Terpstra hypothesis test.

2.2 Mann and Whitney's Expansion on the Kendall Test

In 1947, Henry Mann and D. R. Whitney worked on establishing a test to determine whether one random variable is stochastically larger than the other. This test was formative to the works of both Jonckheere and Terpstra.

First we will state some definitions in line with Terpstra's 1952 paper [10]. Let $X_{k,i}$ be a collection of n independent, continuous random variables for which $1 \leq i \leq m_k$ and

$$\sum_{k=1}^{r} m_k = n$$

holds. the $k^{\underline{th}}$ cluster of random variables, $\cup_{1 \leq i \leq m_k} X_{k,i}$, obeys the same distribution as some variable X_k, being in some sense thought of as observations of X_k. Here, one notes there are r clusters. Many tests have been proposed to determine, from a finite collection of observations of these variables, a null hypothesis that all observations across all clusters are identically distributed, against the alternative hypothesis that there is some monotone trend. First consider the test popularized by Mann and Whitney in their 1947 publication, which evaluates a test statistic for a two-sample problem in terms of a counting question. Namely, consider for $r = 2$ the clusters $X_{1,i}$ and $X_{2,j}$ where $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$ and the hypotheses:

$$H_0 : X_1 \text{ and } X_2 \text{ have equal cumulative distributions.}$$
$$H_a : X_1 \text{ is stochastically smaller than } X_2.$$

A modern formulation of the Mann-Whitney test statistic W is given by

$$W = \sum_{i,j} N_{<}(X_{1,i}, X_{2,j}),$$

where the function $N_{<}$ is defined piecewise by

$$N_{<}(a,b) = \begin{cases} 1 & a < b \\ 0 & \text{elsewhere.} \end{cases}$$

The behavior of the statistic W can be understood by recognizing that it counts the number of pairs that align with the alternative hypothesis H_a. This testing approach is especially suitable when dealing with scenarios where $n_1 \neq n_2$. This method is far more robust than previous techniques for identifying trends, while maintaining strong performance in terms of asymptotic behavior and reliability.

2.3 Jonckheere and Terpstra

In 1952, T. J. Terpstra wrote an essay which establishes a method of testing for trend in clusters [10]. In 1954, A. R. Jonckheere wrote an essay which tackles the same topic [11]. Both mathematicians were credited for the following.

Using the W statistic of Mann and Whitney as presented above, Terpstra then defines a statistic for the $(r > 2)$-sample problem, given in particular for $1 \leq i \leq n_k$ and $1 \leq j \leq n_\ell$ by the summations

$$T = \sum_{k,\ell:\ 1\leq k<\ell\leq r} \left[\sum_{i,j} N_<(X_{k,i}, X_{\ell,j}) \right]. \tag{1.2.1}$$

Moreover, if we let

$$W_{k,\ell} = \sum_{i,j:\ 1\leq i\leq n_k, 1\leq j\leq n_\ell} N_<(X_{k,i}, X_{\ell,j}),$$

then we may write

$$T = \sum_{k,\ell} W_{k,\ell},$$

where k, ℓ are under the same compound inequality impositions as in (1.2.1) Note that the statistic T is nothing but the sum of the Mann-Whitney statistics over all possible pairings of clusters, intending to provide insight as to the behavior of any individual pairing of clusters with respect the alternative hypothesis H_a.

It is important to note that this is not weighted as it takes pairings k, ℓ. That is, the test statistic T does not in this form take into account the gap size $\ell - k$ between clusters. For more discussion regarding weighted methods for for statistics regarding trend, see Kendall's 1938 work, specifically his S- Statistic [9], .

One prominent focus of Terpstra's work is its discussion of various properties of the null distribution T. Terpstra presents the following theorem.

Theorem 2.1 (Stochastic Independence Under H_0). *Let $Y_1, ..., Y_n$ be (i.i.d.) with n-dimensional probability set R_n and for any point $(y_1, ..., y_n) \in R_n$ assign the ranks $r_1, ..., r_n$ by*

$$r_p = \frac{1}{2} \sum_{1 \leq q \leq m} \operatorname{sgn}(y_p - y_q) + \frac{n+1}{2}$$

(so that the ranks also have a probability distribution on the space). For any m such that $1 \leq m \leq n$, an associated partition of the random variables into subsets $Y_1, ..., Y_m$, $Y_{m+1}, ..., Y_n$ and the collections of statistics $\{U\}$ and $\{V\}$ so that $\{U\}$ is a function only of the specific permutation on $r_1, ..., r_m$ and $\{V\}$ is a function only of the specific permutation on $r_{m+1}, ..., r_n$. Then, $\{U\}$ is stochastically independent of $\{V\}$.

This result shows that if H_0 is true, then the Mann-Whitney statistics used as intermediate steps to obtain Terpstra's T-statistic are (completely) independent.

Terpstra then goes on to classify, with proof and in rather broad terms, some families of alternative hypotheses for which the test rejecting H_0 for T above some threshold T_α is consistent. This result is given biconditionally, giving legitimate sense and feel of completeness to this component of Terpstra's analysis of T. Terpstra also shows that the null distribution of the statistic T is asymptotically normal [10], meaning that as the sample size approaches infinity, the distribution converges with the normal distribution.

The prior treatment may seem ill-equipped to handle the instance where at least some pair of observations in the original sample of n observations take on equal values, referred henceforth as a tie. In fact, this is pointed out in a note appended by Jonckheere in his 1954 essay. The problem of ties becomes a problem for ties between clusters [11]. When ties occur within a given cluster, of course, the value of the test statistics is not affected and the validity of the test remains.

Terpstra justifies the original definition of the statistic T by way of the assumption that the sample of observations be independent and continuous [10]. In particular, by the assumption of continuity, one verifies with probability that any finite sample of observations does not have ties.

In practice, however, measurements made to finite precision will permit such ties to occur, including between clusters under some reasonable or natural definition in a particular application. Terpstra thus proposed a remedy to this issue that is coherent in context with the rest of the theoretical analysis performed [10]. Terpstra gives the recommendation that, "for the case, that equal observations occur, this definition may be extended by increasing $[W_{k,\ell}]$ with one half for each pair$[(X_{k,i}, X_{\ell,j})]$ of equal observation" [10]. That is, the recommendation is to use a test statistic called JT given by

$$JT = \sum_{k,\ell:\ 1\leq k<\ell\leq r} \left[\sum_{i,j} N_<(X_{k,i}, X_{\ell,j}) + \frac{1}{2} N_=(X_{k,i}, X_{\ell,j}) \right],$$

for $1 \leq i \leq n_k$ and $1 \leq j \leq n_\ell$. Here, $N_=(a,b)$ takes the value 1 where $a = b$ and 0 otherwise. See immediately via associativity that

$$JT = T + \sum_{k,\ell:\ 1\leq k<\ell\leq r} \left[\sum_{i,j} N_=(X_{k,i}, X_{\ell,j}) \right].$$

Here, we also require $1 \leq i \leq n_k$ and $1 \leq j \leq n_\ell$. In the case that there are no ties between clusters, then consequently see that $JT = T$ because as we execute the summation the quantity $N_=(X_{k,i}, X_{\ell,j})$ is uniformly zero. And, when there are ties between clusters, the recommendation in the literature is to use JT [10, 11]. For the purpose and scope of this project and report, we denote by JT the Jonckheere-Terpstra test statistic, which has desirable properties for a test of trend against a monotone alternative for clusters of varying sizes in which cross-cluster ties are permitted.

3 Conclusion

3.1 The Results

	Austin	Corpus Christi	Dallas
0.065	JT = 216345 P-Value = 0	JT = 113410 P-Value = 0	JT = 521375 P-Value = 0
0.07	JT = 86784 P-Value = 0.0271	JT = 40434 P-Value = 3.828e-05	JT = 232511 P-Value = 0
0.075	JT = 28840 P-Value = 0.3741	JT = 14290 P-Value = 0.0001	JT = 104746 P-Value = 6.114e-09
0.08	JT = 7511 P-Value = 0.0206	JT = 3716 P-Value = 1.095e-05	JT = 43674 P-Value = 0.0057
	El Paso	Houston	San Antonio
0.065	JT = 149694 P-Value = 1.60e-09	JT = 470048 P-Value = 0	JT = 214366 P-Value = 2.324e-06
0.07	JT = 42566 P-Value = 4.15e-08	JT = 262190 P-Value = 0	JT = 90464 P-Value = 0.0786
0.075	JT = 12898 P-Value = 5.24e-07	JT = 140042 P-Value = 7.09e-09	JT = 30911 P-Value = 0.1065
0.08	JT = 3815.5 P-Value = 0.0183	JT = 74821 P-Value = 0.0009	JT = 10484 P-Value = 0.6513

Table 1: Note that values below 1.00e-10 have been reported as approximated zero. Green p-values indicate rejecting the null (i.e. a decreasing trend is identified), while red p-values indicate accepting the null (i.e. a decreasing trend is not identified).The metropolitan areas have been abbreviated to the first city named in the area for the sake of readability.

The goal of this research was to identify decreasing trends in ground-level ozone pollution in many of the largest metropolitan areas in Texas. The null hypothesis states that no trend was identified and the alternative hypothesis states that a decreasing trend was identified. A conservative significance level of $\alpha = 0.01$ was chosen, in order to determine with more certainty the presence of a decreasing trend. For three out of the six metropolitan areas, a decreasing trend was identified at all four thresholds tested, see Table 1. Those areas were Corpus Christi, the Dallas-Fort Worth-Arlington area, and the Houston-The Woodlands-Sugar Land area. In El Paso, a decreasing trend was identified at all but the highest threshold, though the $p - value$ at the threshold of 0.08 for El Paso was 0.0183, which is incredibly close to our significance level. Thus, I would state with confidence that a decreasing trend is present in El Paso. For the Austin-Round Rock area and the San Antonio New Braunfels area, a decreasing trend was identified only at the lowest threshold, and the $p - values$ for the three higher thresholds for both areas were significantly higher than the significance level of $\alpha = 0.01$, implying that there is likely not a statistically significant decreasing trend present in either area. It is important to note that the the size of the $p - value$ is not an indicator of the magnitude of the trend identified. Instead, it is merely an indicator of the presence of a trend. That is, a very small p-value does not indicate that over time the ground-level ozone levels dropped steeply, but instead that they dropped steadily over time. It does appear that ground-level ozone level have been decreasing over time in major metropolitan areas in the last 43 years. However, in spite of this result, there is not evidence to state any relationship between this decrease and legislation.

3.2 Discussion

There is much room for further analysis and study on this topic, including more interesting uses for the Jonckheere-Terpstra test statistic, further analysis of the raw data provided, and continued exploration of interesting patterns that were noticed in the data.

It would be pertinent to consider new methods for performing the Jonckheere-Terpstra hypothesis test on the data. There is only one R-Studio package that performs the Jonckheere-Terpstra hypothesis test, with little supporting documentation. Unfortunately, because of time and ability restrictions, creating a test from scratch was not feasible.

There are many questions remaining, specifically regarding the data acquired from the EPA. For each of the six regions analyzed in this study, there were almost no readings for December 31st of any year, with no clear reason as to why. Additionally, further analysis of the reason for large chunks of missing data may provide insight into the behavior of ground-level pollution readings during those time periods.

In the research conducted for this analysis, there was some discussion of methods to use the Jonckheere-Terpstra test statistic as an indicator of extremity of decline. However, these methods are outside the bounds of my current knowledge and ability. There are other patterns in ground-level ozone readings that I think would be interesting to consider. I hypothesize that coastal regions have lower overall ground-level ozone readings than regions located further from the coast. Further analysis of this topic and other related topics would be interesting. As a whole, there is significant room for continuing research and discussion regarding this topic.

4 Acknowledgments

A special thanks to Dr. Kaitlin Hill and St. Mary's University. Preliminary research efforts were funded by NSA grant number H98230-23-1-0017 and NSF grant DMS-2244093 to Dr. Javier Rojo through the RUSIS@IU program.

5 Appendix

5.1 A: Images

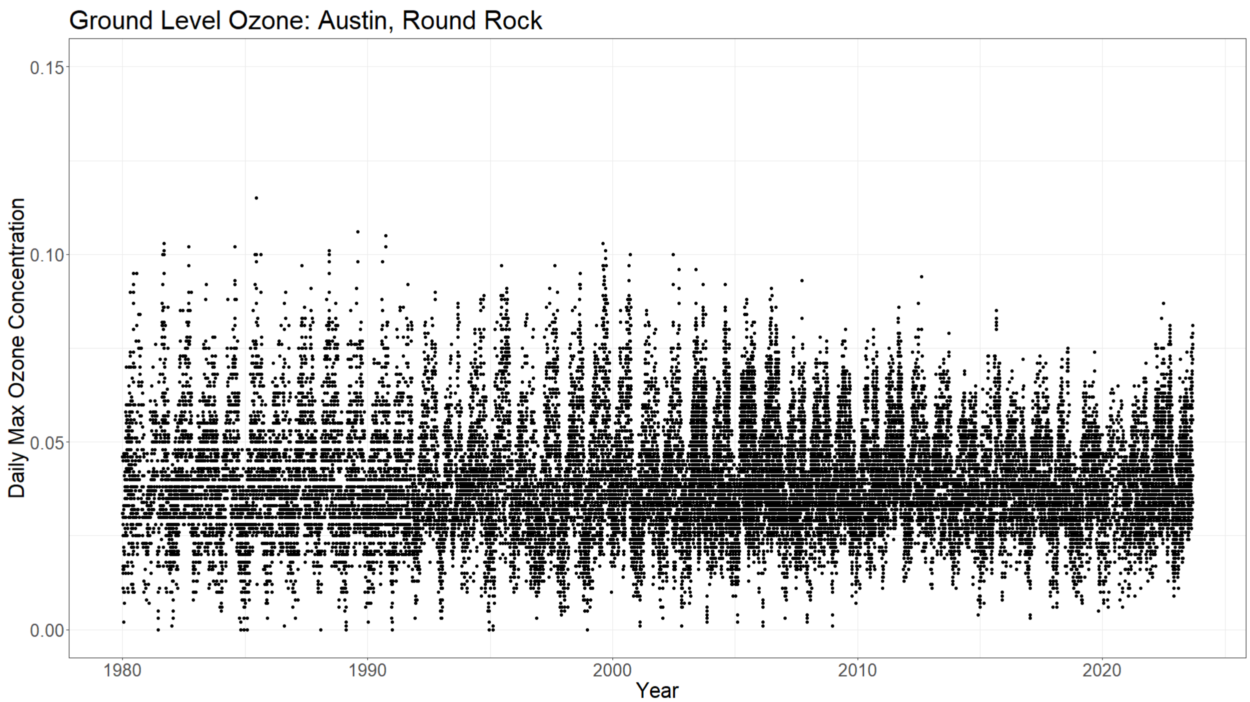

Figure 3: This is a scatter plot of the raw data from the EPA for the Austin-Round Rock area, from 1-1-1980 to 9-4-2023. Note that the white lines in this image and the following scatter plots from 1980 to 1990 are caused by rounding error, which was no longer an issue post 1990 due to improvements in technology.

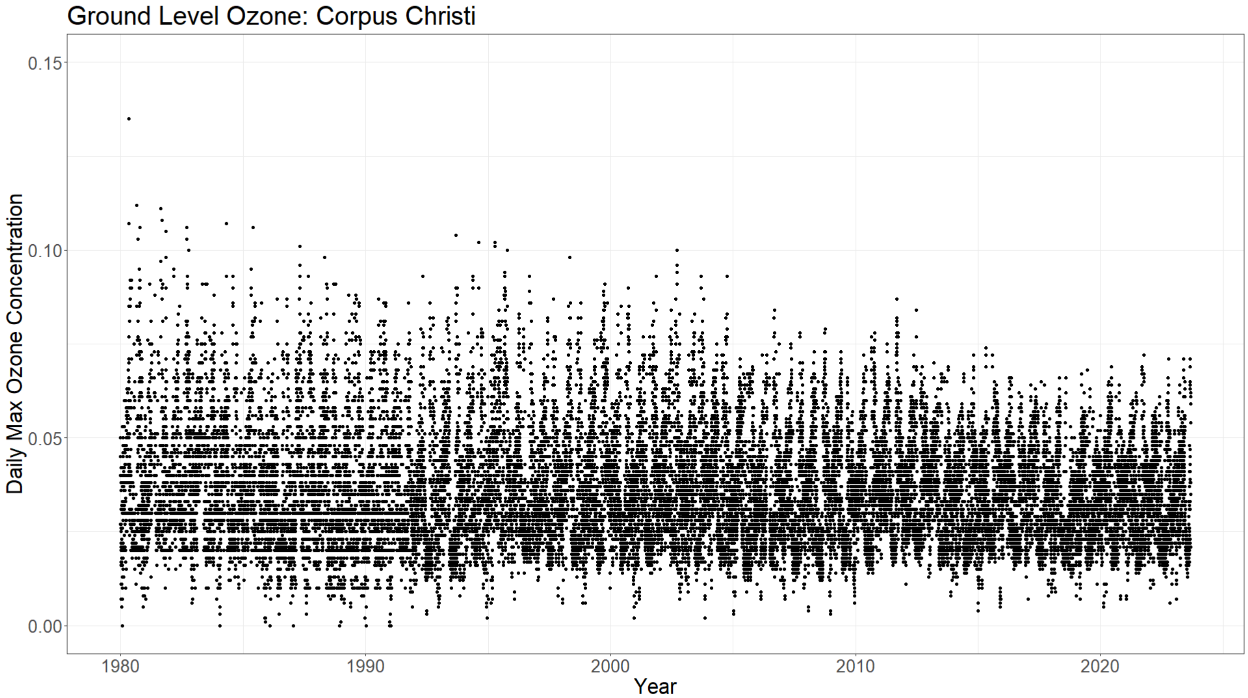

Figure 4: This is a scatter plot of the raw data from the EPA for Corpus Christi, from 1-1-1980 to 9-4-2023.

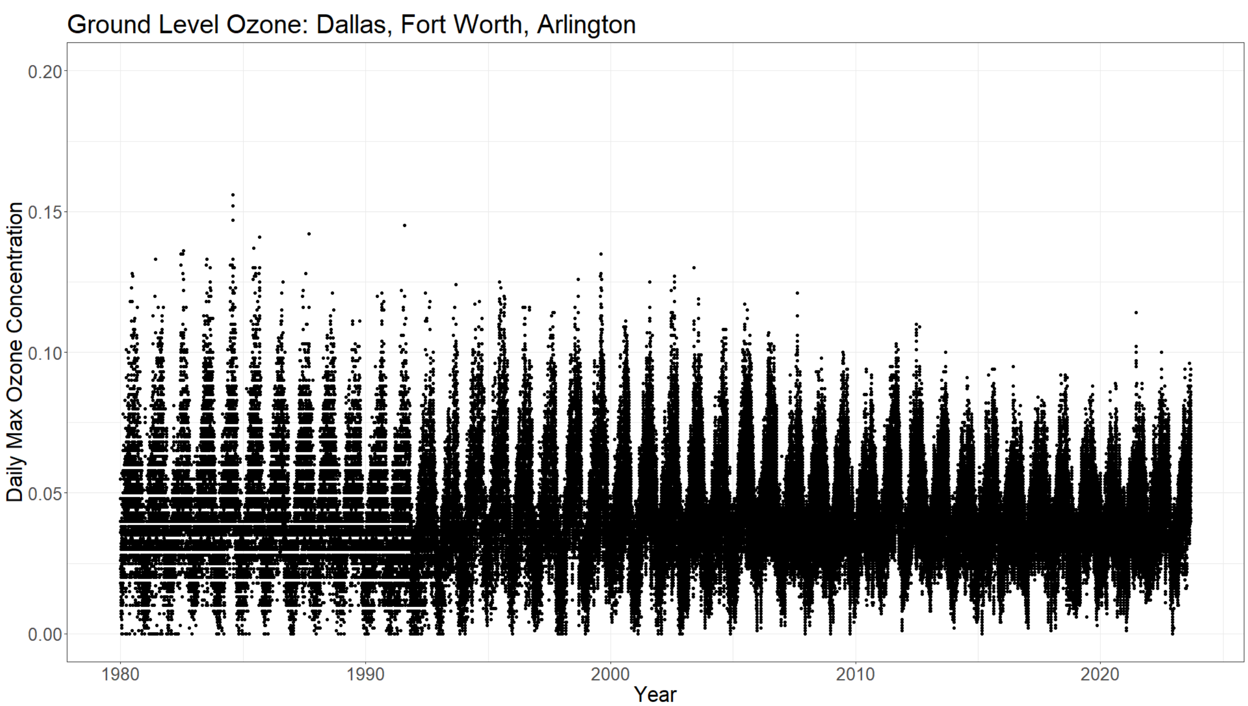

Figure 5: This is a scatter plot of the raw data from the EPA for the Dallas-Fort Worth-Arlington area, from 1-1-1980 to 9-4-2023.

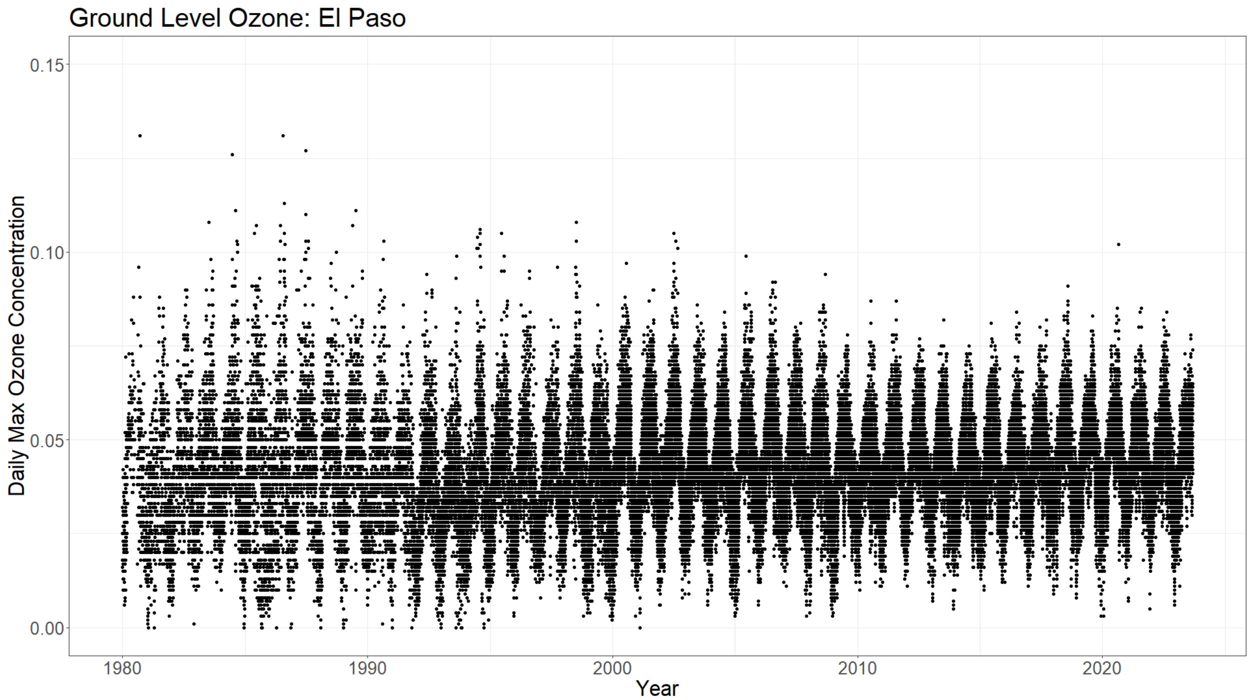

Figure 6: This is a scatter plot of the raw data from the EPA for El Paso, from 1-1-1980 to 9-4-2023.

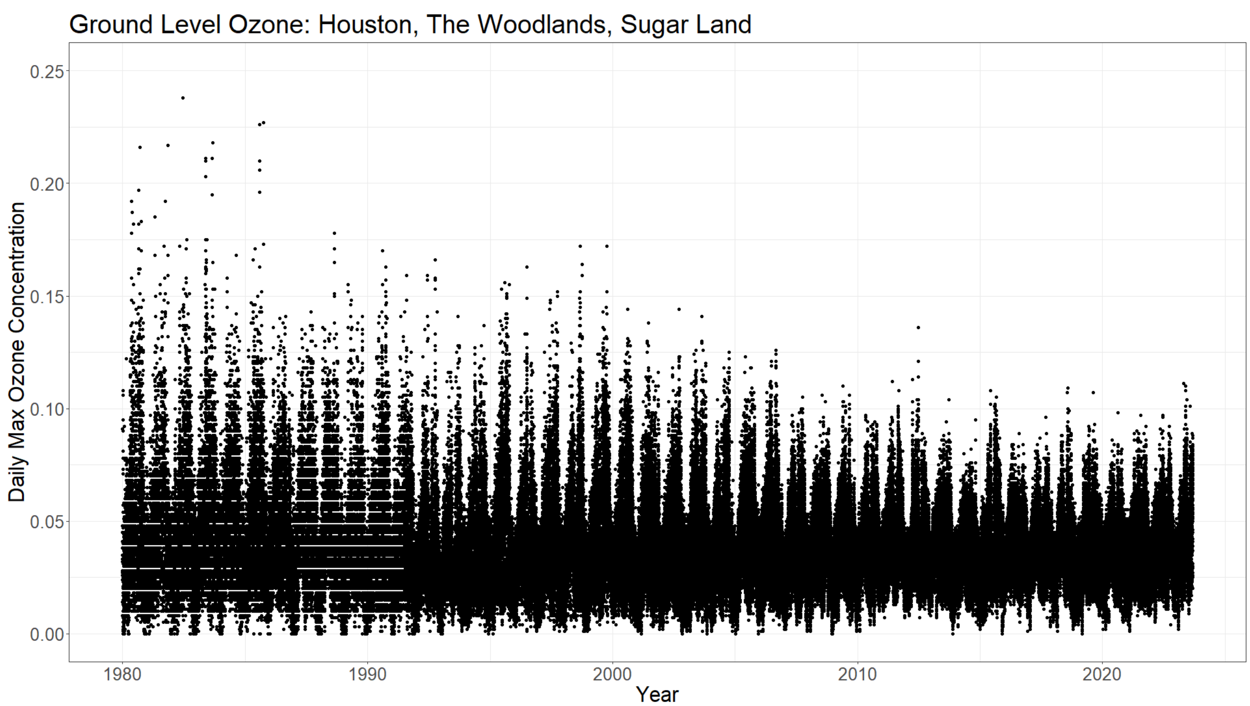

Figure 7: This is a scatter plot of the raw data from the EPA for the Houston-The Woodlands-Sugar Land area, from 1-1-1980 to 9-4-2023.

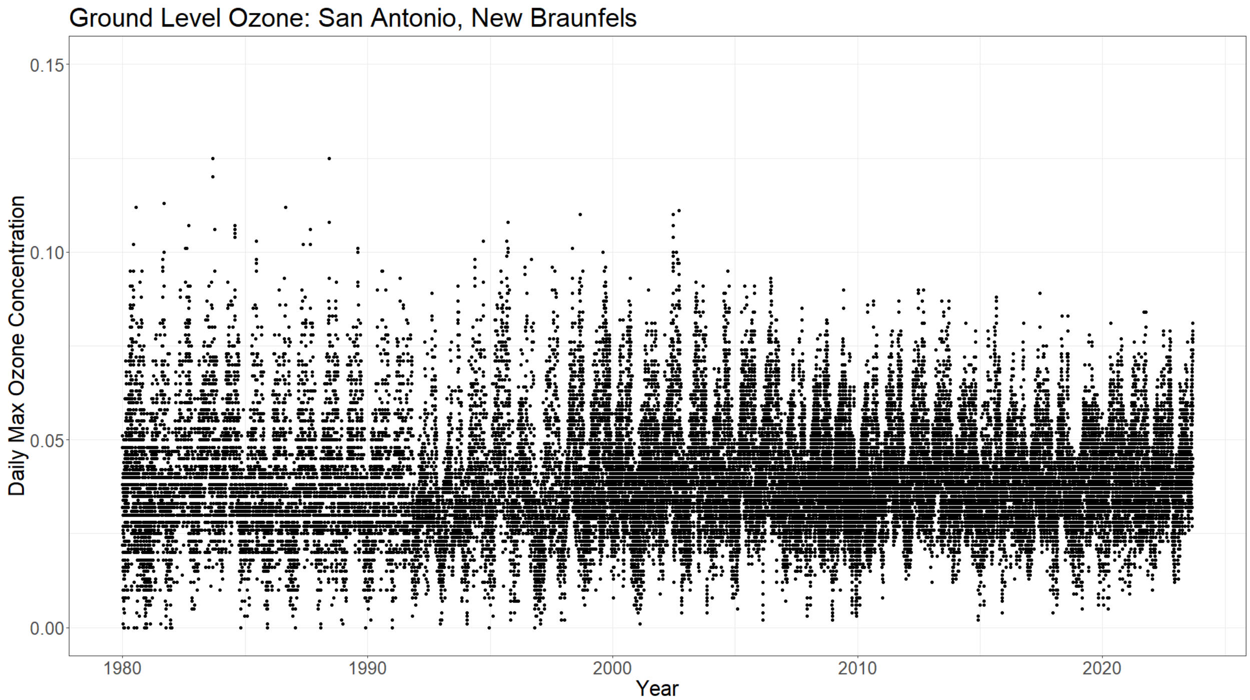

Figure 8: This is a scatter plot of the raw data from the EPA for the San Antonio-New Braunfels area, from 1-1-1980 to 9-4-2023.

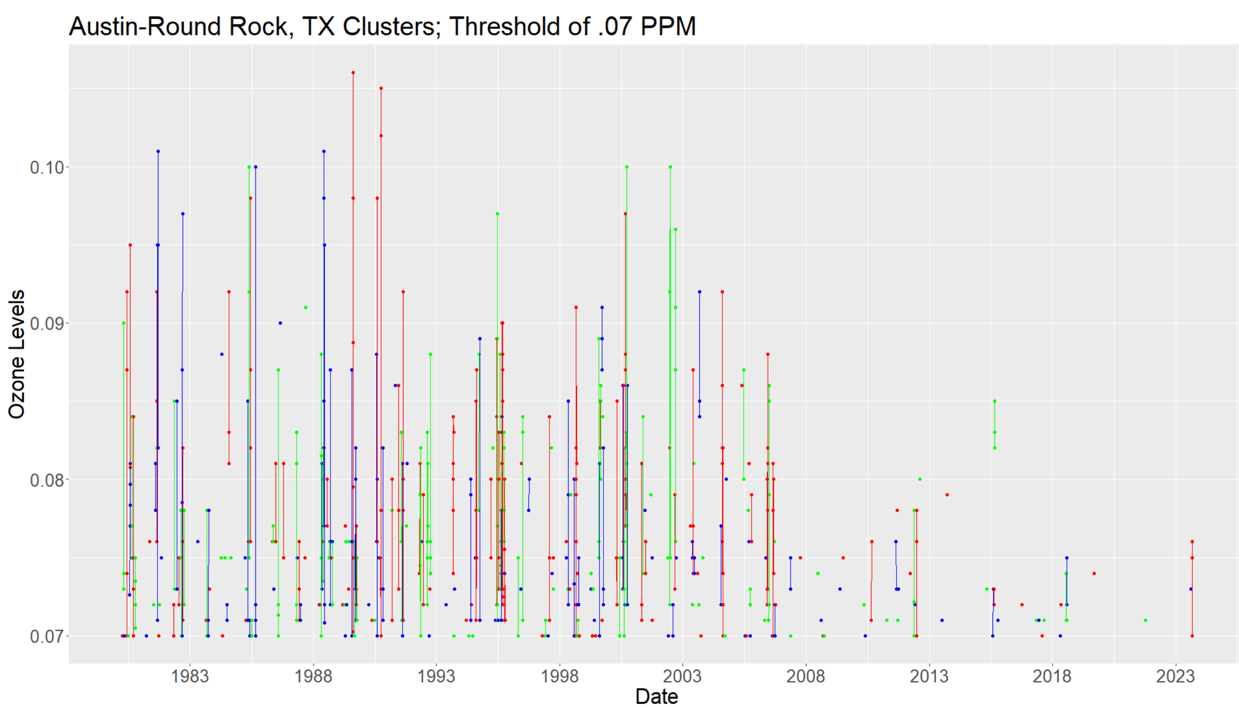

Figure 9: This is the Austin-Round Rock area Clusters above a threshold of 0.07. The clusters rotate though the colors red, green, and blue.

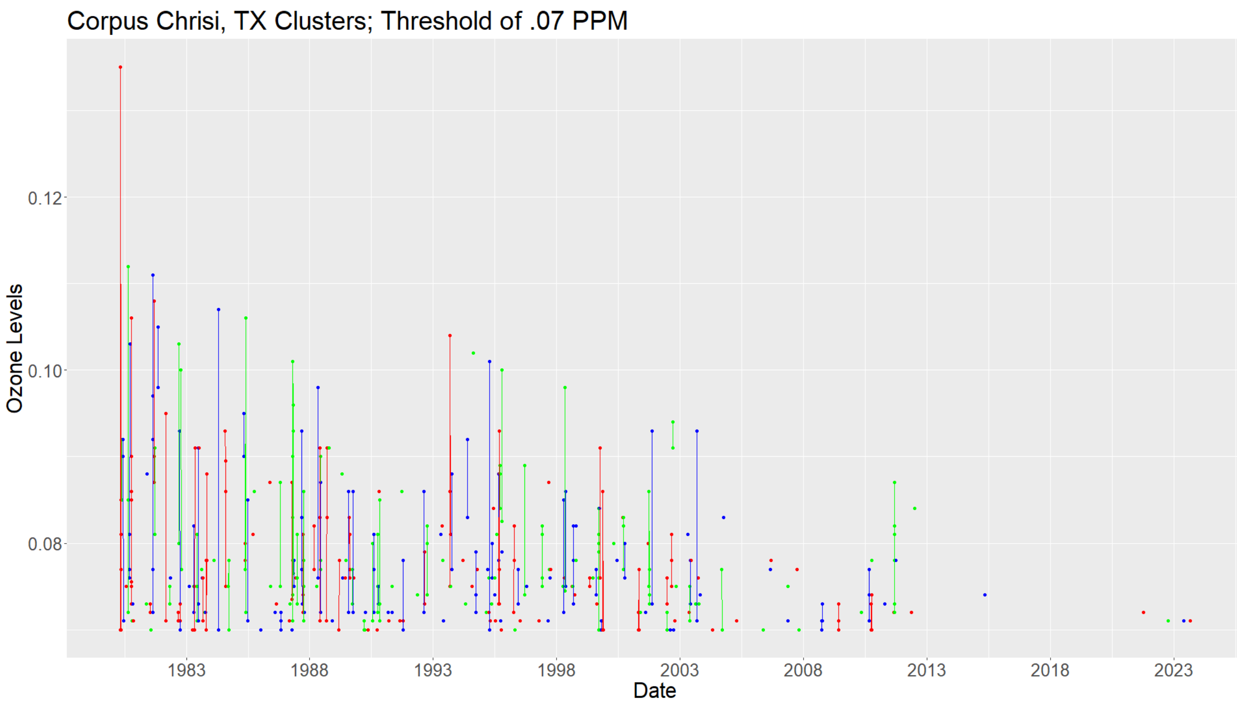

Figure 10: This is the Corpus Christi Clusters above a threshold of 0.07. The clusters rotate though the colors red, green, and blue.

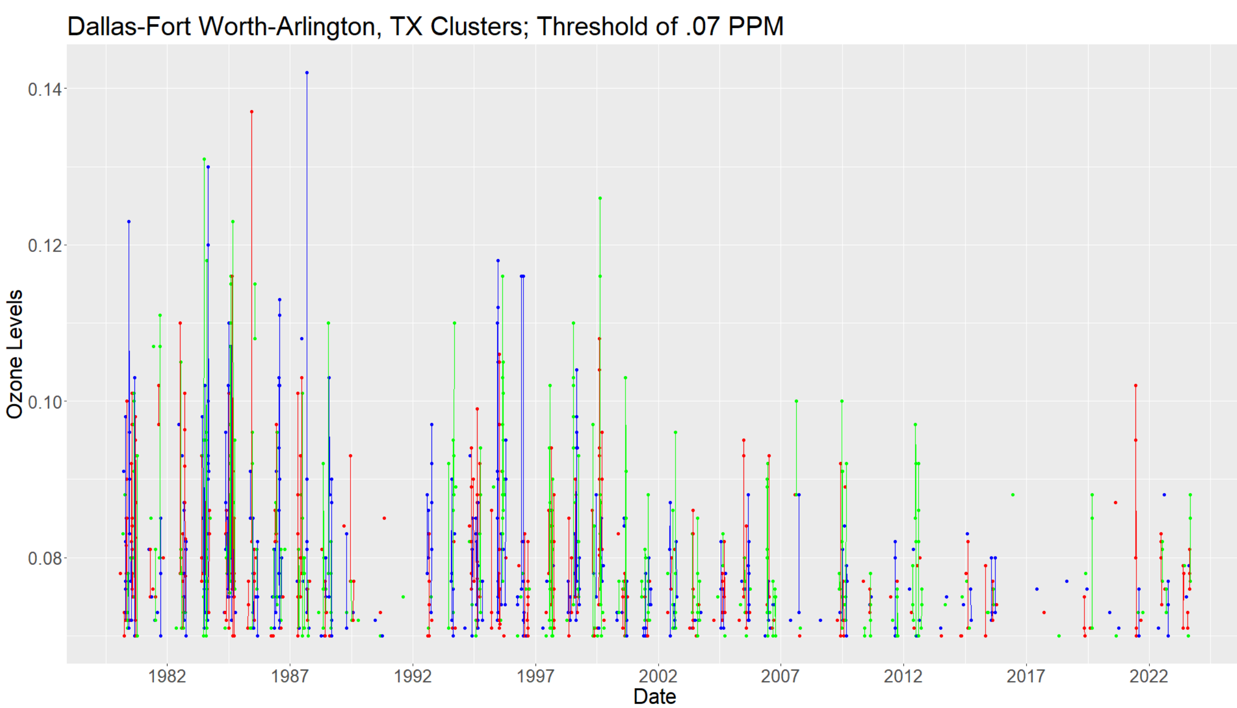

Figure 11: This is the Dallas-Fort Worth-Arlington area Clusters above a threshold of 0.07. The clusters rotate though the colors red, green, and blue.

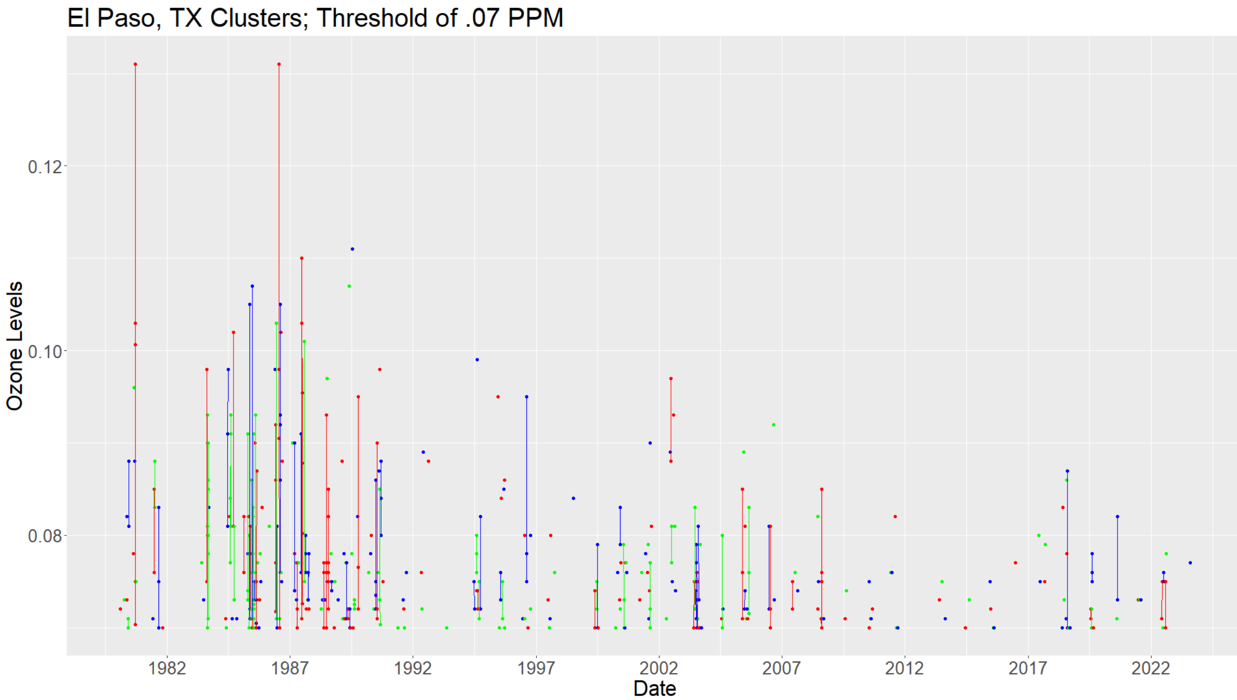

Figure 12: This is the El Paso Clusters above a threshold of 0.07. The clusters rotate though the colors red, green, and blue.

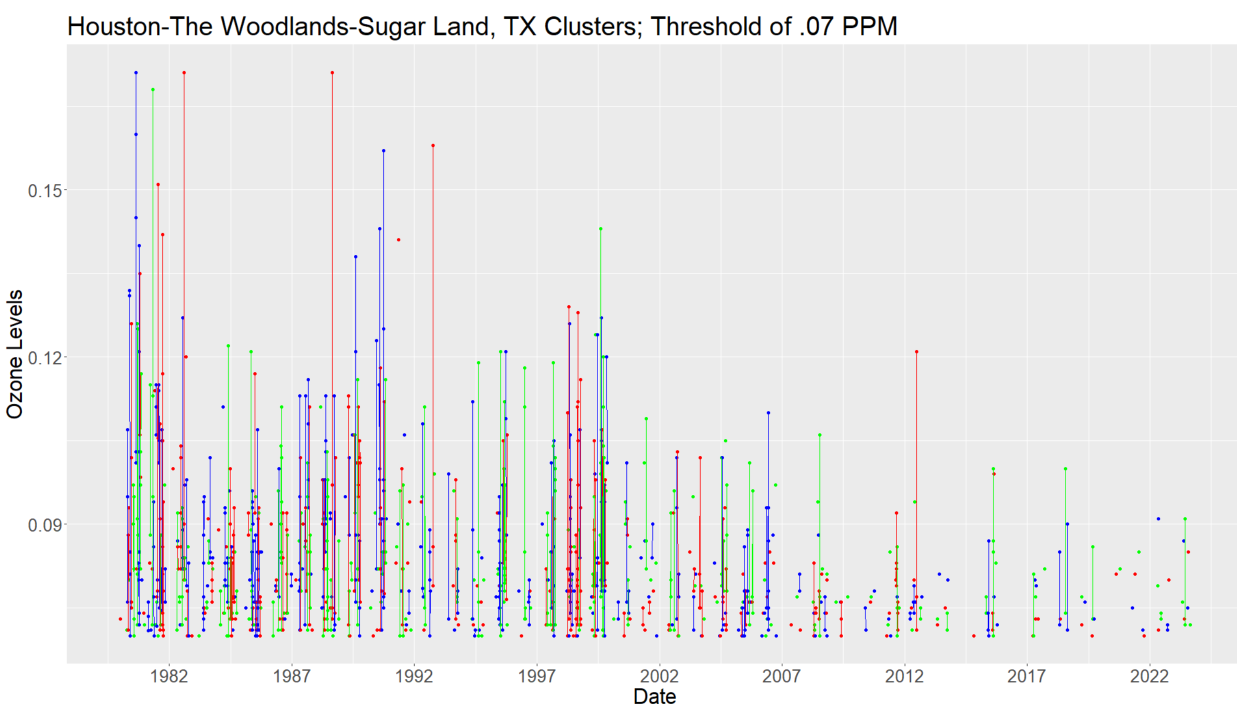

Figure 13: This is the Houston-The Woodlands-Sugar Land area Clusters above a threshold of 0.07. The clusters rotate though the colors red, green, and blue.

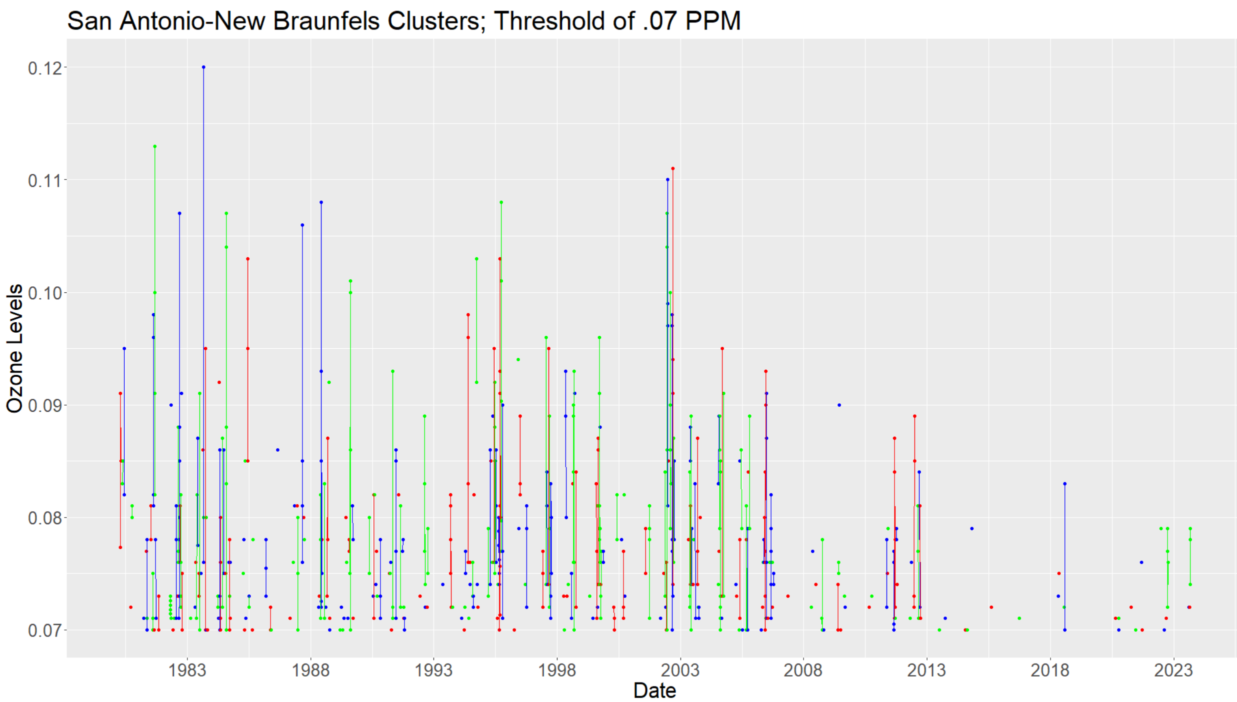

Figure 14: This is the San Antonio-New Braunfels area Clusters above a threshold of 0.07. The clusters rotate though the colors red, green, and blue.

5.2 B: Code

```
##This is the code which establishes the clusters.
if (!require("pacman")){
  install.packages("pacman", "parallel", "parallelly")
}
pacman::p_load("readxl")

loadingdock <- function(){
  workbookpath <- readline(prompt = "enter site filepath :")
  #get the path we need
  sitesheetidentifiers <- excel_sheets(workbookpath)
  #get the sheets at the named path
  sdvsheets <- lapply(sitesheetidentifiers, function(site)
  read_excel(workbookpath, sheet = site, na = "NA"))
  #read the sheets into a resultant list
  for (site in sdvsheets){ #for each site with an sdv in the result
    site$Date <- as.Date(site$Date, format = "%m/%d/%Y")
    #sanitize the dates
    site$Date <- as.Date(site$Date) #remove timestamps
    site$Daily.Max.8.hour.Ozone.Concentration <-
    as.numeric(site$Daily.Max.8.hour.Ozone.Concentration)
    #sanitize the air reading
    site <- as.data.frame(site)
    #solves problem of reverse UTC coercion later
  }
  return(sdvsheets)
```

```
  #return the data in its natural format, where appropriate, as a list of sdv
}

getoutdir <- function(){
  outdir <- readline(prompt = "enter output file directory :")
  return(outdir)
}

getthreshold <- function(){
  threshstr <- readline(prompt = "enter numeric threshold :")
  result <- as.numeric(threshstr)
  return(result)
}

makecopy <- function(obj){
  result <- obj
  return(result)
}

traverseforzeroes <- function(iterable){
  result <- c()
  lasttwo <- c(0, 0)
  for (jndex in 3:(length(iterable)-1)){
    lasttwo[1] <- iterable[jndex-2]
    lasttwo[2] <- iterable[jndex-1]
    if (lasttwo[1] == 0 && lasttwo[2] == 0 && iterable[jndex] == 0
    && iterable[jndex+1] != 0){
```

```
      result <- append(result, jndex)
    }
  }
  return(result)
}

countfrombin <- function(bin, jndex){
  result <- 0
  for (item in bin){
    if (item < jndex){
      result <- result + 1
    }
  }
  return(result)
}

assemblerowassigngroup <- function(rowid, sdv, groupbin){
  currentrow <- sdv[rowid,]
  if (currentrow[1,"Exceedances"] == 0){
    currentrowaugment <- makecopy(currentrow)
    currentrowaugment$Group <- c(0)
  }
  else{
    currentrowaugment <- makecopy(currentrow)
    groupcount <- countfrombin(groupbin, rowid)
    currentrowaugment$Group <- c(groupcount)
  }
```

```
  return(currentrowaugment)
}

main <- function(){
  sdvlist <- loadingdock()
  outdir <- getoutdir()
  thresh <- getthreshold()
  for (sdv in sdvlist){
    site <- ifelse("Site.Id" %in% colnames(sdv), sdv[1,"Site.Id"],
    "site_unnamed")
    base::message("Processing site ",site, " for clustering at threshold "
    ,thresh)
    sdv$Exceedances <- ifelse(sdv$Daily.Max.8.hour.Ozone.Concentration <
    thresh, 0, 1)
    groupbinlibrary <- traverseforzeroes(sdv$Exceedances)
    destructablecopy <- makecopy(sdv)
    mutablewithzeroes <- destructablecopy[0,]
    for (rowid in 1:length(sdv$Date)){
      mutablewithzeroes <- rbind(mutablewithzeroes,assemblerowassigngroup
      (rowid, sdv, groupbinlibrary))
    }
    mutablewithzeroes <- mutablewithzeroes[order(mutablewithzeroes$Date),]
    base::message("mutable with zeroes prepped for site ",site)
    wzfilename <- paste(outdir, site, "_clusters_with_zeroes.csv", sep = "")
    write.csv(mutablewithzeroes, file = wzfilename)
    base::message("mutable with zeroes; write successful")
    mutablenozeroes <- makecopy(mutablewithzeroes)[0,]
```

```
    for (rowid in 1:length(mutablewithzeroes$Date)){
      testrow <- mutablewithzeroes[rowid,]
      if (testrow[1,"Exceedances"] != 0){
        mutablenozeroes <- rbind(mutablenozeroes, testrow)
      }
    }
    base::message("mutable without zeroes prepped for site ",site,"
    from prior mutable")
    nzfilename <- paste(outdir, site, "_clusters_no_zeroes.csv", sep = "")
    write.csv(mutablenozeroes, file = nzfilename)
    base::message("mutable without zeroes; write successful")
  }
}

main()

##This is the code which uses both linear interpolation and ARIMA Model.
install.packages("forcast")
install.packages("imputeTS")
# Load necessary libraries
library(forecast)
library(imputeTS)
#Reads the data you wish to interpolate.
AllData <- read.csv("Insert Reference CSV File")
#Pulls only the ozone readings, excluding unnessecary data.
Data <- AllData$Daily.Max.8.hour.Ozone.Concentration
#Linearly interpolate missing data
```

```
IData <- na_interpolation(Data)
#a placeholder for where the final interpolation
w/ ARIMA model and linear interp.#
SIData <- IData
# Create a time series object
TSData <- ts(IData, frequency = 365)
# Adjust the frequency according to the seasonality of your data

NAData <- is.na(Data)
x <- 1
while (x < length(NAData)) {
  if (NAData[x]) {
    n <- 1
    while (x + n < length(NAData) && NAData[x + n]) {
      n <- n + 1
    }
    if (n >= 10) {
      SARIMA <- auto.arima(TSData[1:x], seasonal = TRUE)
      forecast_values <- forecast(SARIMA, h = n)
      SIData[x:(x + n - 1)] <- forecast_values$mean
    }
  }
  x <- x + n
}
write.csv(SIData, file = "File Name", row.names = TRUE)
write.csv(IData, file = "File Name", row.names = TRUE)
```

```
#This is the code which performs the JT Test
#download and load in DescTools Package
install.packages("DescTools")
library(DescTools)
#Test File
Test<- read.csv("Test File.csv")
JonckheereTerpstraTest(Reading~Cluster, Test, alternative = "decreasing")
```

References

WHO. (2022). Billions of people still breathe unhealthy air: New who data. https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data

Ozone. (2023). https://www.lung.org/clean-air/outdoors/what-makes-air-unhealthy/ozone#:~:text=Long%2Dterm%20ozone%20exposure%20is,main%20driver%20of%20total%20mortality.

WHO. (2015). *Reducing global health risks: Through mitigation of short-lived climate pollutants* (tech. rep.). World Health Organization. Retrieved December 5, 2023, from http://www.jstor.org/stable/resrep33063

Smith, R. L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science*.

Houston metropolitan statistical area profile. (2021). https://www.houston.org/houston-data/houston-metropolitan-statistical-area-profile

Economic development. (2023). https://www.elpasotexas.gov/economic-development/economic-snapshot/snapshot-overview/

Us census bureau. (2022). https://www.census.gov/quickfacts/fact/table/corpuschristicitytexas/POP060210

Ground-level ozone guidelines. (2023). https://www3.epa.gov/region1/airquality/index.html#:~:text=On%20October%201%2C%202015%2C%20EPA,in%20the%20presence%20of%20sunlight.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*.

Terpstra, T. J. (1952). The asymptotic normality and consistency of kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae (Proceedings)*.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Box, G. E. P., & Jenkins, G. M. (1968). Some recent advances in forecasting and control. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *17*(2), 91–109. Retrieved November 21, 2023, from http://www.jstor.org/stable/2985674

Mann, H. (1945). Nonparametric tests against trend, *Econometrica*.

Mann, H., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Math. Stat.*