Spring 2023

A Change-Point Analysis of Air Pollution Levels in Silao, Mexico and Fresno, California

Rachael Goodwin

A Change-Point Analysis of Air Pollution Levels in Silao, Mexico and Fresno, California

Rachael Goodwin

June 2023

1 Introduction

In this project, we analyzed PM10 levels from 2010 to 2019 in the city of Silao, Guanajuato, Mexico as well as PM2.5 and PM10 levels in the San Joaquin Valley. PM10 refers to matter less than 10 micrometers in diameter, including smoke and soot particles, and is used as a measure of air quality. It can be generated by natural causes as well as emissions from factories and other industries. High PM10 levels are linked to conditions such as asthma, bronchitis, and strokes. Thus, determining whether the mean level of PM10 has changed over time is important to evaluate potential environmental health risks. These pollutants are known to adversely affect agriculture and the tourism industry, as people may be less likely to travel to places known for poor air quality. In places with a large emphasis on the tourism industry, this could be disastrous to the economy and create further negative effects for the region.

The primary purpose of this analysis is to determine if either location experienced an adverse shift in pollution due to the negative effects associated with high levels of air pollution. Because of the numerous health and economic risks correlated with increases in air pollution levels, understanding what factors contributed to the rise in pollution is helpful for mitigating future increases. The methods used within this paper can be generalized to other locations with significant air pollution, and will provide useful information as to when shifts in pollution levels could have occurred, which can then be cross referenced with urban and agricultural developments, and weather data dating to the time of the identified change point to identify potential causes associated with the rise in pollution.

To model the air pollution levels, we assumed that there was a seasonal component to the data and that a stationary autoregressive moving average process of orders $(1, 1)$, commonly known as ARMA(1,1), would be suitable for modeling the data. These assumptions were supported by examining the autocorrelation and partial autocorrelation function (ACF and PACF) plots. Normality of the residuals was also assumed, and confirmed based on the results of a Shapiro-Wilk test.

In this paper, we first focus on the data gathering and the methods used to analyze it before discussing the results and analyzing them, as well as potential further avenues of research or significance of the results. Section 2 will focus on the significance of the data and the locations chosen while Section 3 is dedicated to the methods employed in this analysis. Section 4 reports on the results of the research and Section 5 discusses potential reasons for any changes observed in the time series analysis.

2 Background

We analyzed PM10 data from two cities: Fresno, CA and Silao, Guanajuato, Mexico. The San Joaquin data were obtained from the Fresno Supersite, maintained by the National Aeronautics and Space Administration (NASA), while the Silao data was obtained from the Instituto Nacional de Ecología y Cambio Climático (Secretaría de Medio Ambiente y Ordenamiento Territorial, 2021), which measures levels of various air pollutants in Mexico. The data sets were chosen due to their location in regions with high levels of air pollution. The San Joaquin Valley specifically is known for its poor air quality; in particular, it continues to have extremely high ozone levels despite successful efforts to improve the air quality. Ozone is a pollutant known to reduce lung function, and thus its high levels within the San Joaquin Valley represent a danger to the local population's health (EPA, 2000). The air quality within the region failed to meet federal standards in 2001, causing the U.S. Environmental Protection Agency (EPA) to downgrade its rating to severe, given the high level of the ozone within the area (EPA, 2001). The city of Fresno, California, is associated with particularly poor air quality, due to its location in the San Joaquin Valley as well as human activity such as wood burning and vehicular emissions. This region has a particularly high rate of asthma related emergency room visits and hospitalizations, and there was a correlation between higher rates of such events and high ozone and particulate pollution levels (Meng et al., 2010).

Potential causes for the high PM2.5 levels may be related to emissions: particulate nitrate, associated with soil emissions contributes heavily to such pollution in winter months (Ying et al., 2008). As soil emissions increase, predicted PM2.5 levels were seen to increase as well (Kleeman et al., 2019). Reducing oxides of nitrogen was found to be effective in controlling PM2.5 levels. However efforts to reduce sulfur oxides have minimal effect (Chen et al., 2014). Additionally while reducing oxides of nitrogen can provide some decreases in PM2.5 levels, such gains are often offset by summer forest fires and residential wood burning in the winter, meaning that cooler temperatures may provide an increase in

PM2.5 that cannot be reduced by focusing on soil emissions. Moreover, fog is associated with the production of ammonium nitrate as well, which also contributes to PM2.5 emissions (de Foy et al., 2019).

With regards to the PM10 levels within the San Joaquin Valley, agricultural activities were found to have a significant impact. For instance, fungal spores, plant detritus, and pollen grain were around 11-15% of PM10 mass in the fall of 2000 (Chow et al., 2015). As with the PM2.5 index, PM10 is also highly correlated with soil emissions, specifically the sand or clay content of the soil (Carvacho et al., 2001).

In comparison, the poor air quality within Silao may be due to human activity rather than primarily resulting from natural phenomena and disasters. Specifically, there may be a correlation with the construction of automobile factories near the city; at least several hundred have been built within the region of Guanajuato within the past 10 years. The higher levels of pollution may be a result of increased emissions from these factories. Other potential causes include the expansion of Silao's brickwork industry, whose factories burn oil and plastic, possibly contributing to regional air pollution.

3 Exploratory Data Analysis

Our initial goal in examining the data was to determine an adequate model, and to determine if there may be correlation or seasonality within the residuals, as well as a potential mean shift in the data. In the analysis of the Silao data, which our model was initially developed for, we gathered observations measuring PM10 levels every hour in Silao from 2010 to 2021. The 2020 data set had a large

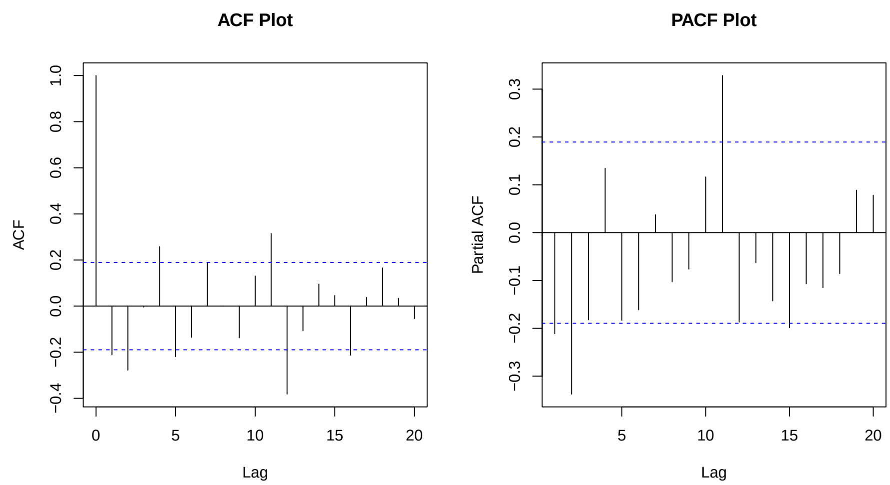

Figure 1: These graphs represent the ACF and PACF of the differenced monthly air pollution measurements for the Silao data.

amount of missing observations and the data for 2021 was not fully collected at the time this analysis was begun; as such, only observations from 2010 to 2019 were included. The PM10 data for each month were averaged to obtain a total of 120 observations. Upon initial examination of the time series data, it seemed there was a fairly significant change around $t = 80$ in the pollution levels that could be a potential mean shift, as well as potential seasonality within the data with both annual and semiannual components. Thus, when analyzing the residuals, we first attempted to difference the data at lag 1 and lag 12, i.e., subtracting the $t-1$ observation from the t-th observation and then subtracting the $(t-12)$-th observation from the t-th observation to remove seasonality. While observing the autocorrelation function (ACF) plot of these residuals in Figure 1, several lags were significant, indicating that the residuals were not independent.

Given its ineffectiveness for this data set, instead of relying on differencing to obtain satisfactory residuals, we next modeled the seasonality and subtracted it from the raw data to create independent and identically distributed (iid) residuals. To visualize the changes in PM10 over time, we first focused on modeling the seasonal component of a logarithmic transformation of the data. When examining the ACF plot of the transformed Silao data in 4, it appeared to have high correlation at lags 6 and 12, indicating that the data likely did contain a seasonal component, and that the residuals were likely correlated. Upon examination of the ACF plots of these residuals, the ARMA(1,1) process was found to be a satisfactory model for the data.

Based on this initial analysis, the data process consists of a seasonal component, a potential mean shift, and stationary residuals. The seasonal component can be modeled using sine and cosine functions with annual and semiannual periods:

$$\beta_1 \sin(\tfrac{\pi t}{6}) + \beta_2 \cos(\tfrac{\pi t}{6}) + \beta_3 \sin(\tfrac{\pi t}{3}) + \beta_4 \cos(\tfrac{\pi t}{3}).$$

The model created resulted in approximately normal residuals (6) when using the p-value obtained by using the Shapiro-Wilk test for normality, 0.6063, and based on the Q-Q plot for normality, they appear fairly normal.

4 Methodology

Our primary goal was to identify if a mean shift had occurred, which was complicated by the potential lack of iid residuals. Typically, change point analysis assumes that the data has iid residuals and does not have a seasonal trend. Thus, the first step was determining if the residuals were stationary. A step function was modeled based on the data and change-point analysis was utilized to determine whether this change was a statistically significant result. The to-

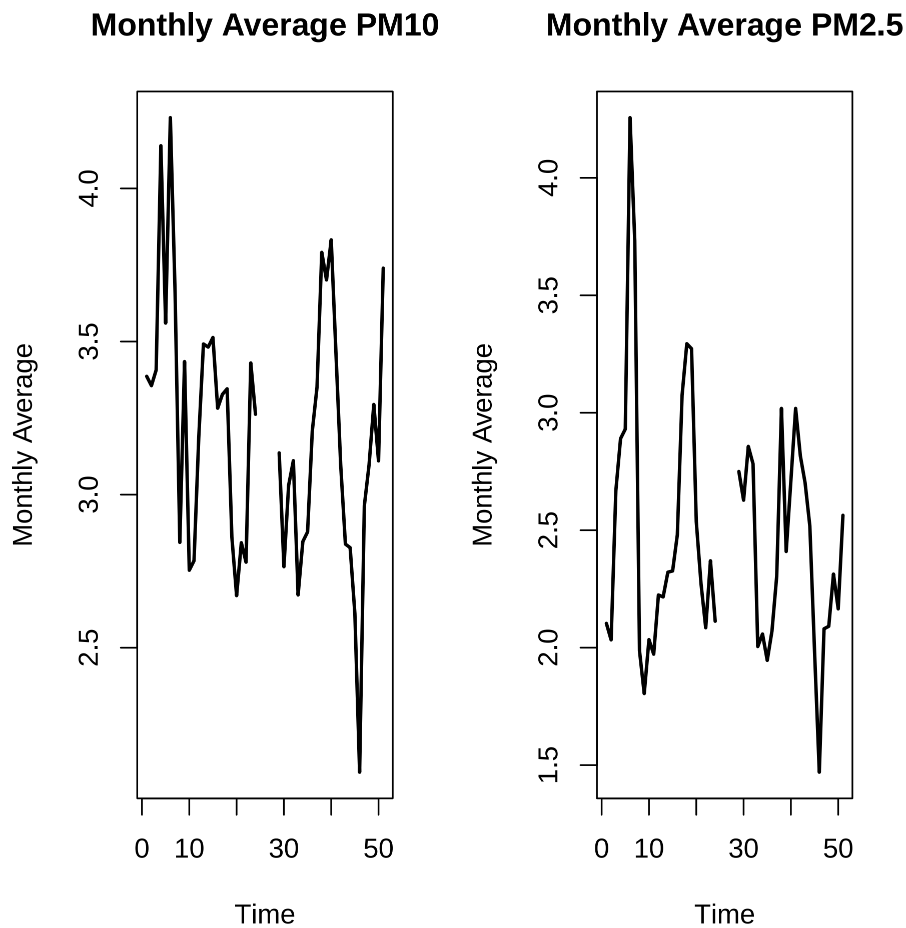

Figure 2: These graphs are time series plots of the log transformed monthly PM10 and PM2.5 levels in Fresno.

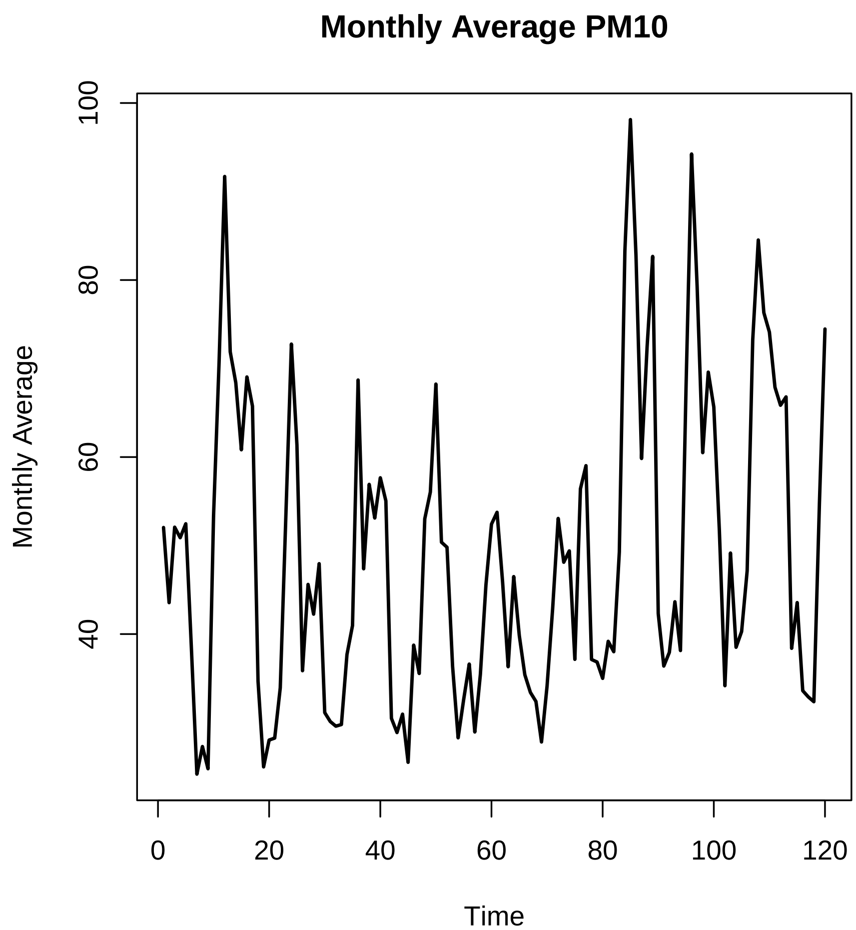

Figure 3: These graphs are time series plots of the monthly PM10 levels in Silao.

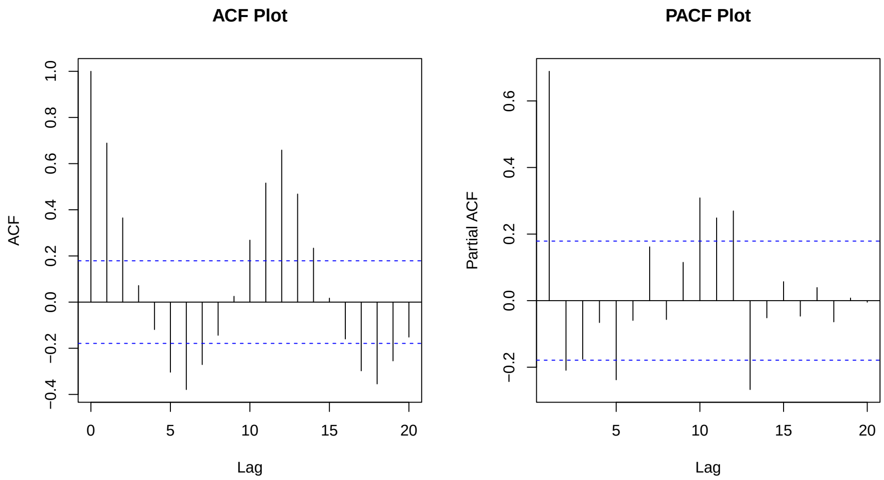

Figure 4: These graphs are the ACF and PACF of the log-transformed Silao data.

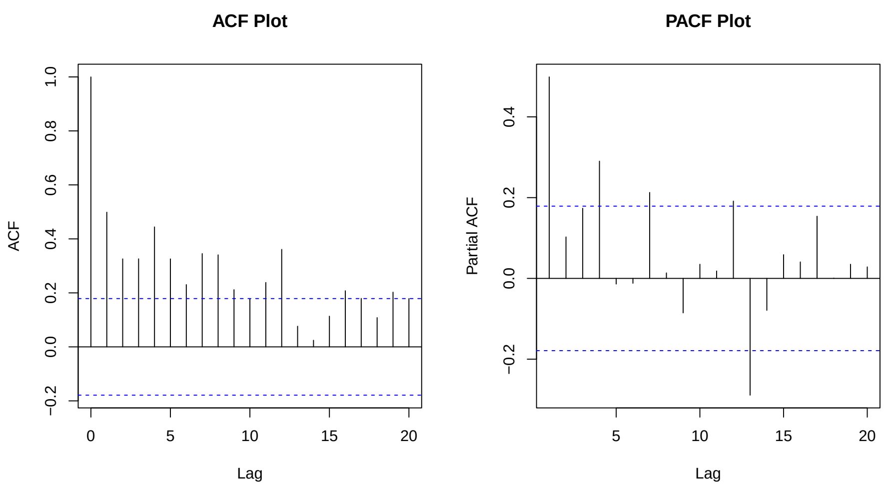

Figure 5: These graphs present the ACF and PACF of the log transformed Silao data with seasonality removed.

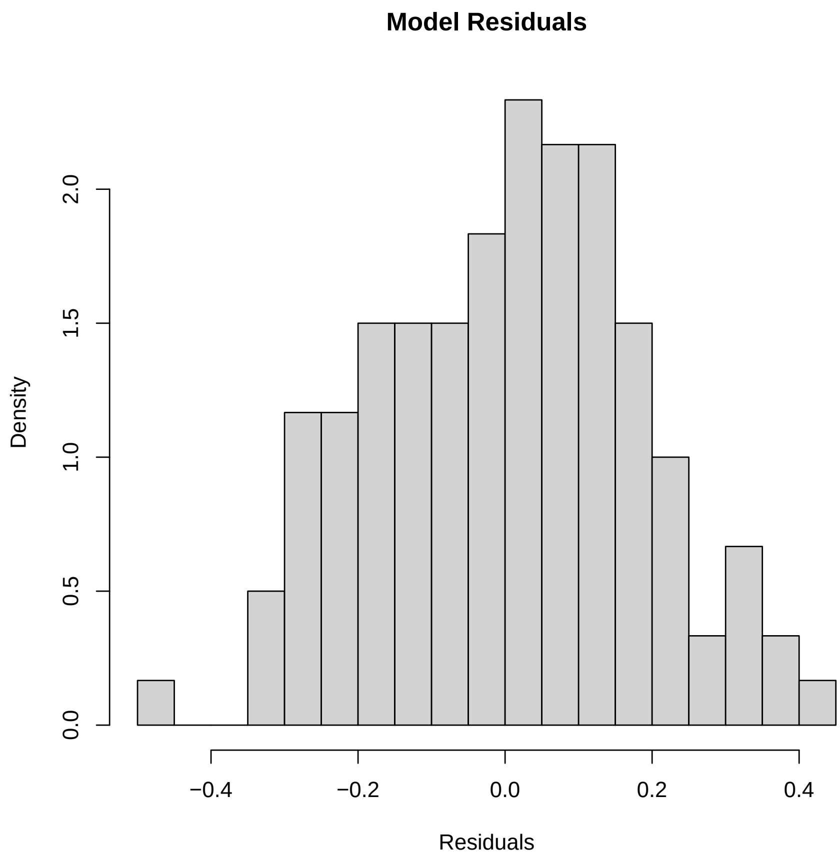

Figure 6: This graph presents a histogram of the residuals of the ARMA residuals and it is approximately normal, further supporting the normality of the data.

tal sum of squares (SST) of step functions that stopped before and after each observation, as well as the overall data, were collected. Then, the SST of the step functions before and after each t value were subtracted from the overall SST, and the results were stored in a vector to be used as test statistics. The minimum value of the SST was then determined to have occurred at $t = 76$ for this data set.

To determine if this result was statistically significant, the next step was to simulate 1000 ARMA(1,1) models as estimates of the residuals. To determine the statistical significance of i_0, the change at $t = 76$, $B = 1000$ bootstrap samples were created. To create such samples, we added each of the simulated 1000 ARMA(1,1) processes with normal innovations to the estimated seasonal component. The seasonality equation determined from the residuals was added to each model separately, and the SST was evaluated for step functions that started and ended at each value of t within every model and then the calculated values were subtracted from the overall SST of each model and stored in a vector of test statistics. The change point test statistic was obtained by computing

$$T_{i_0} = \min_{i \in S} T_i,$$

where

$$T_i = (\mathrm{SST}_{1,i} + \mathrm{SST}_{2,i})/\mathrm{SST}_0.$$

Here,

$$\mathrm{SST}_{j,i} = \sum_{i=1}^{n_j} (y_{j,i} - \bar{y}_j)^2,$$

where $y_{j,i}$ is the i-th observation in the j-th part, $j = 1, 2$, and

$$\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{j,i}}{n_j},$$

where n_j is the number of observations in the j-th part. Thus $\bar{y}_j$ refers to the mean of all observations in the j-th part.

We omitted the first and last 10 observations so that $S = \{11, 12, \ldots, 110\}$. Similarly,

$$\mathrm{SST}_0 = \sum_{j=1}^{2} \sum_{i=1}^{n_j} (y_{j,i} - \bar{y})^2,$$

where

$$\bar{y} = \frac{\sum_{j=1}^{2} \sum_{i=1}^{n_j} y_{j,i}}{n},$$

where n is the total amount of observations in the data set. The index i_0 is the estimated change point, corresponding to the value that minimizes T_i, $i \in S$. The graph of T_j, where $j = i - 10$, is present in 9.

The location of the minimum value was subsequently determined using the which.min command and each minimum location was stored within a separate vector. Each statistic greater than the value of the minimum change point was then averaged to find the p-value, which resulted in a value of 0, a highly significant p-value. Based on this value, we rejected the null hypothesis that such a change could have occurred by random chance.

The same model was utilized for the Fresno PM2.5 and PM10 data due to the similarities between it and the Silao data. The results were similar, although the residuals appear to be uncorrelated in the ACF and PACF, provided below, instead of following the ARMA(1,1) process. Several months had completely missing data so the ACF and PACF are presented in two parts in 7 and 8. Within these data, no such change point was detected, but the low sample size may have contributed to the lack of the algorithm detecting such a change.

5 Results

This section discusses the results of the analysis for both the Silao and Fresno data, as well as how the test statistics were computed. It begins with a discussion of the methods in the Silao data, and then analyzes the Fresno data in comparison to the Silao data, comparing the characteristics they have in common that make the analysis possible.

5.1 Silao

9 indicates that $i_0 = 66 + 10 = 76$ is the estimated change point. **??** represents the calculated monthly PM10 levels, with lines added representing the average PM10 measurement before and after the identified potential change point.

For each of the bootstrap samples, we computed the change point test statistic. Let T_{i_b}, $b = 1, \ldots, B$, denote the test statistic for the b-th bootstrap sample. Then, the p-value of T_{i_0} is computed by $\#\{T_{i_b} < T_{i_0}\}/B$. The resulting p-value was < 0.01, indicating that the **potential change point identified is statistically significant**.

To further our analysis, we split the data before and after this change point to determine if there were other potential change points, but none were statistically significant.

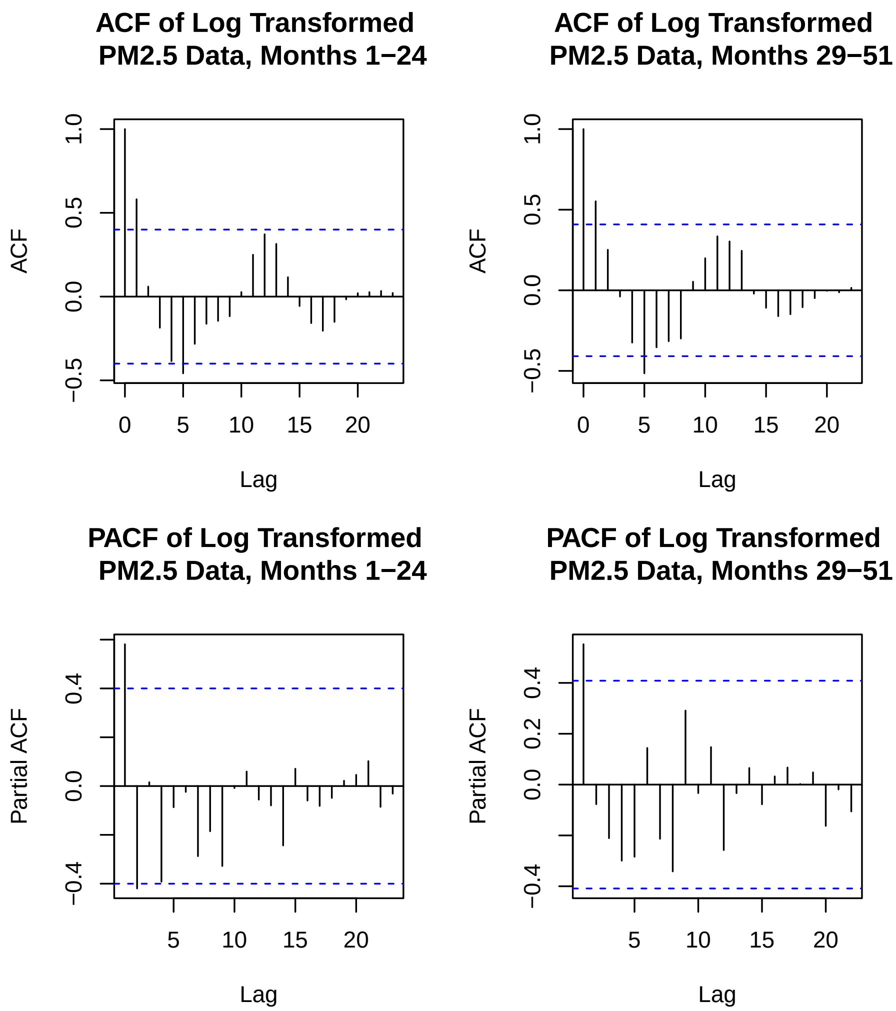

Figure 7: These graphs present the ACF and PACF of the log transformed Fresno PM2.5 data, divided into sections before and after the missing observations.

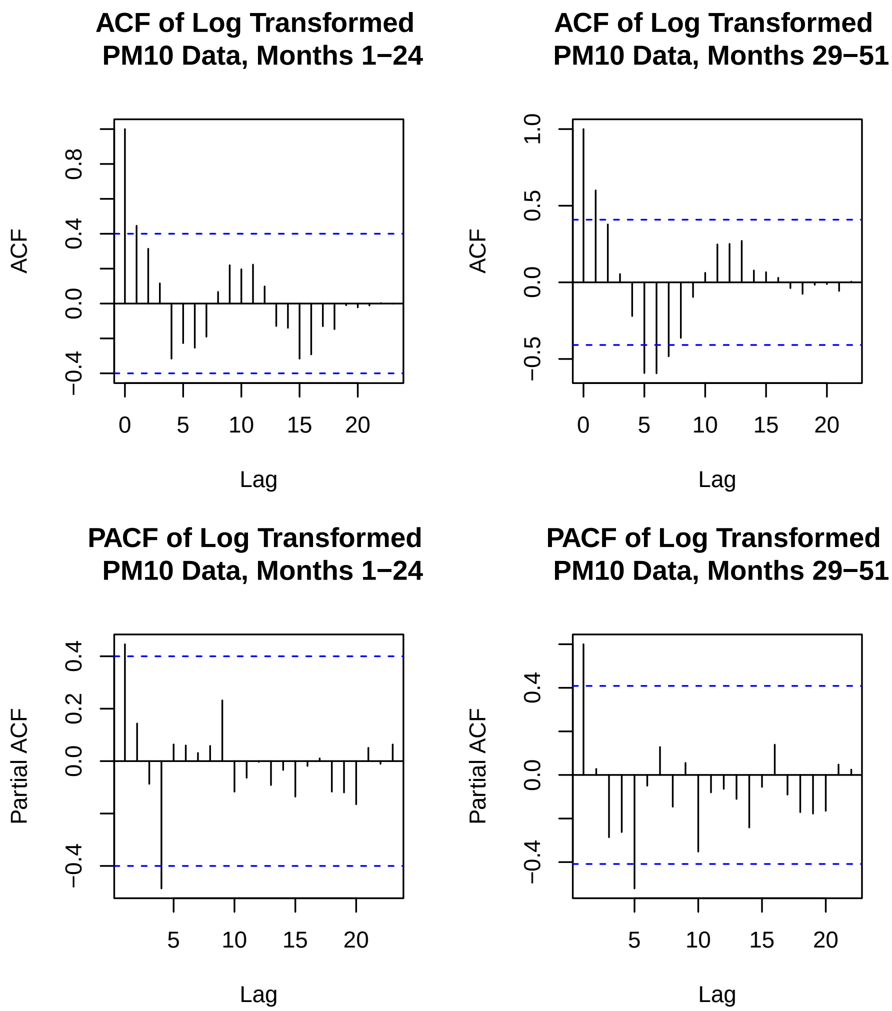

Figure 8: These graphs present the ACF and PACF of the log transformed Fresno PM10 data, divided into sections before and after the missing observations.

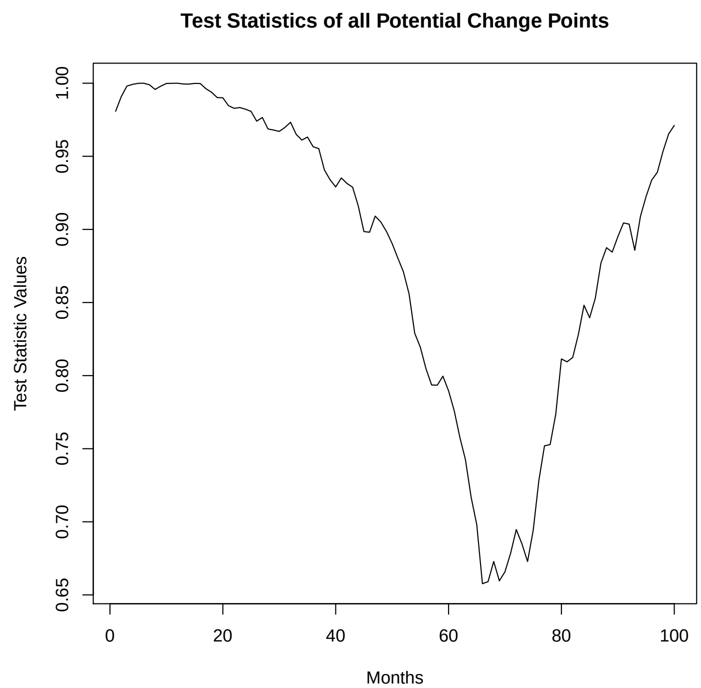

Figure 9: This graph presents the test statistic values of all potential change points within the Silao data.

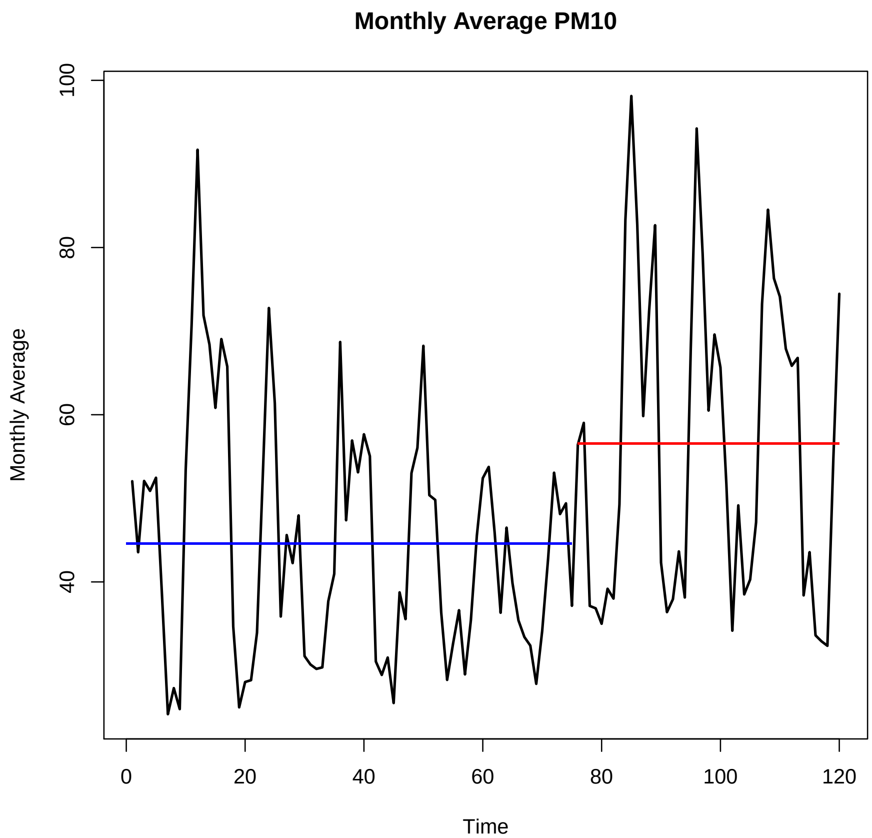

Figure 10: This graph presents the average PM10 levels before and after the estimated change point for the Silao PM10 data.

5.2 Fresno

A similar method was utilized to analyze the data in Fresno given the similarities in the two datasets, such as the presence of a seasonal trend and the use of an ARMA(1,1) process to model the data. Despite the lack of statistically significant evidence, the graphs do demonstrate a downward trend in pollution levels, so there may have been a change that occurred.

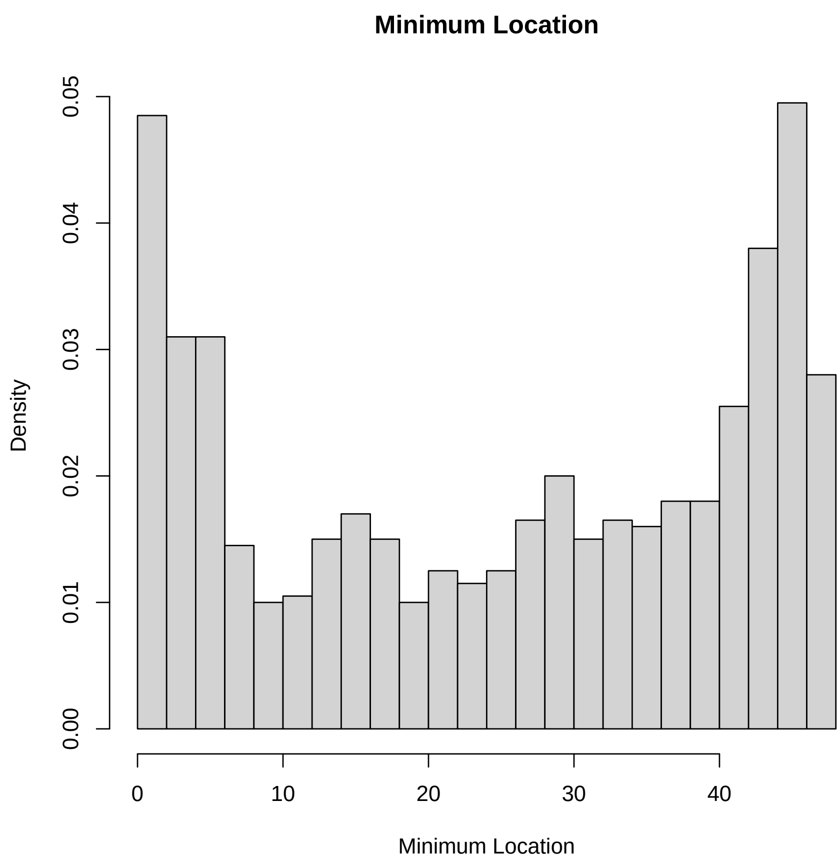

Figure 11: This graph displays the minimum location detected for all iterations of the algorithm for the Fresno PM2.5 data.

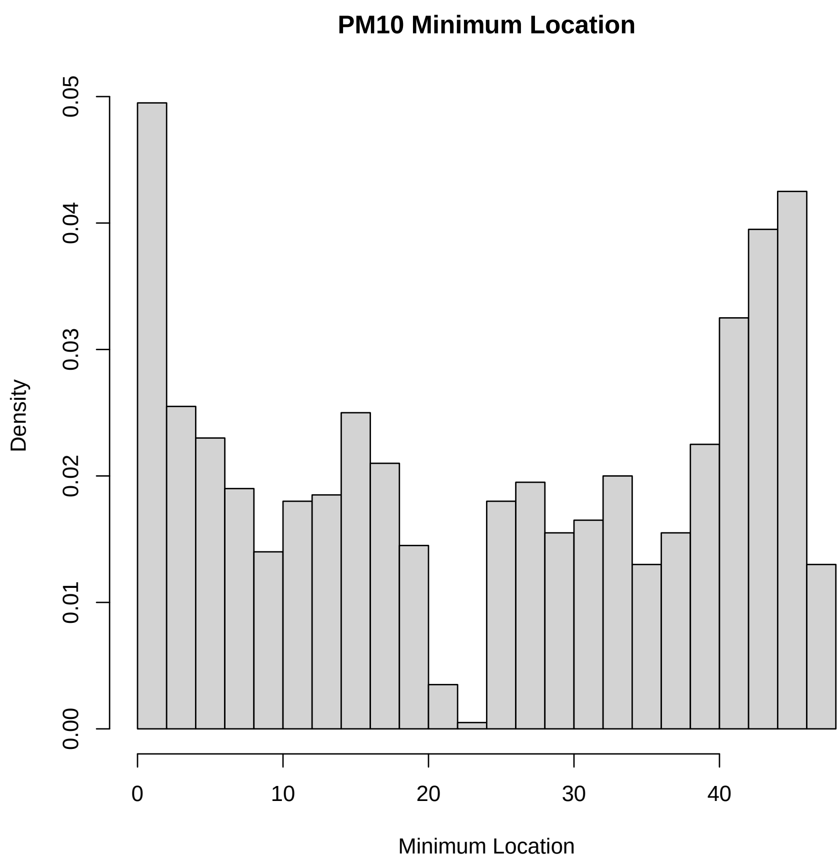

Figure 12: This graph displays the minimum location detected for all iterations of the algorithm for the Fresno PM10 data.

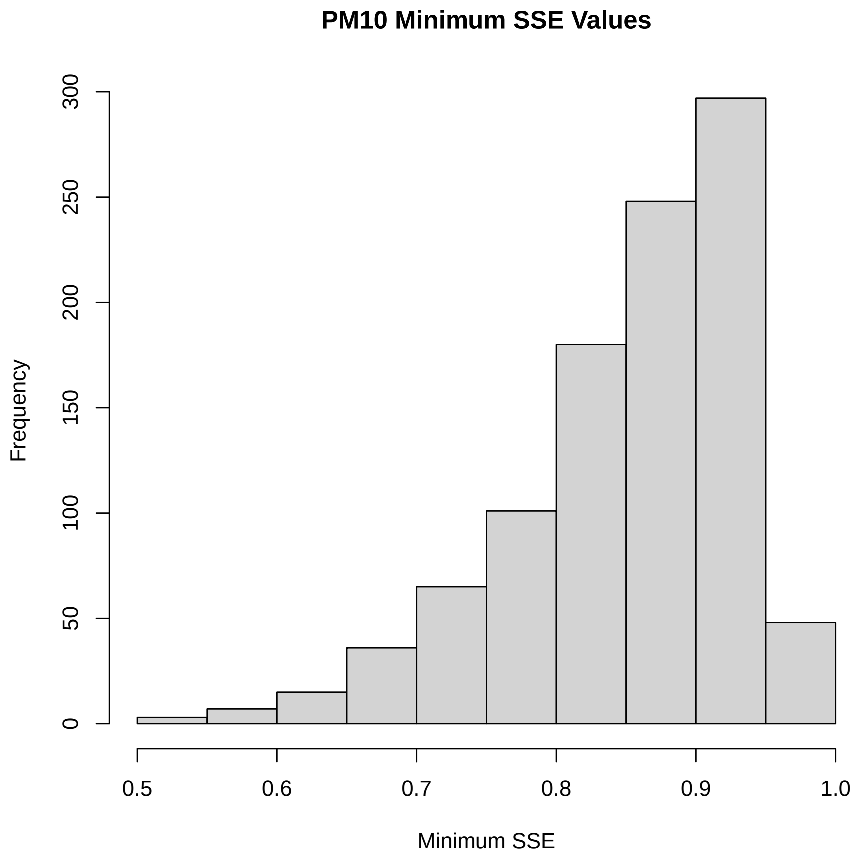

Figure 13: This graph displays the minimum SSE detected for all iterations of the algorithm for the Fresno PM10 data.

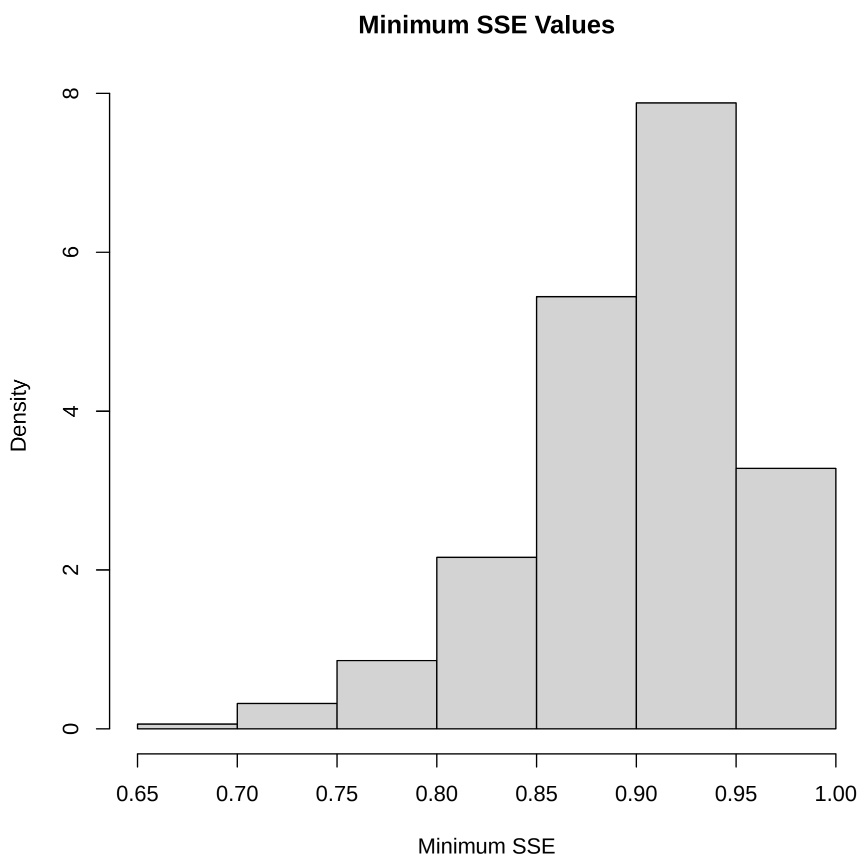

Figure 14: This graph displays the minimum SSE detected for all iterations of the algorithm for the Fresno PM2.5 data.

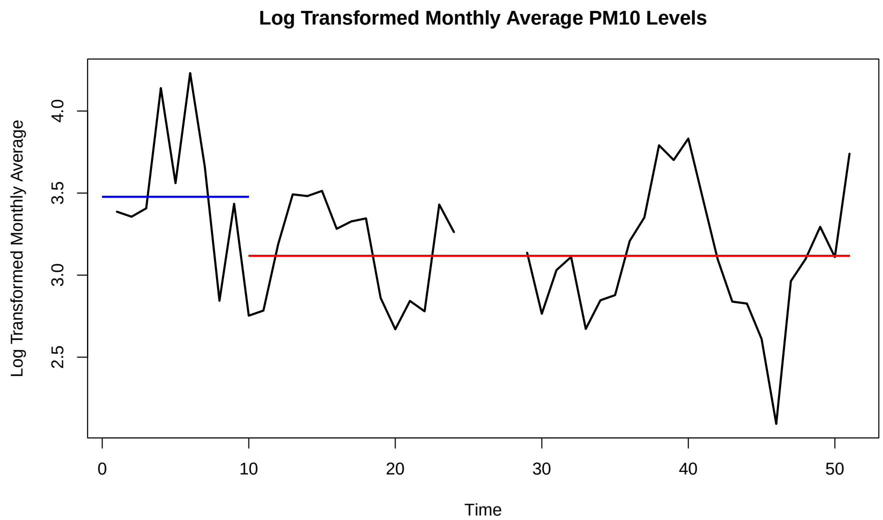

Figure 15: This graph overlays the average PM10 levels before and after the estimated change point on the log transformed monthly pollution data from Fresno.

6 Discussion

In the Silao data, we identified a statistically significant mean shift in the PM10 levels in the 2010–2019 data at the 76th month, corresponding to mid-2016. The increase in the PM10 levels may be linked to the expansion of the automobile industry. In the state of Guanajuato, at least several hundred new factories were built over the past ten years, which could be linked to the increased PM10 levels in Silao (Manufacturing in Guanajuato). Another possible link to the increase in PM10 levels in Silao could be the expansion of the brickwork industry. These factories burn materials such as oil and plastic, generating high amounts of air pollutants (Silao es la sexta ciudad con más contaminación en México).

The Fresno data appeared to have a mean shift in PM10 and PM2.5 levels, but neither was statistically significant. One possible reason for the decrease, even

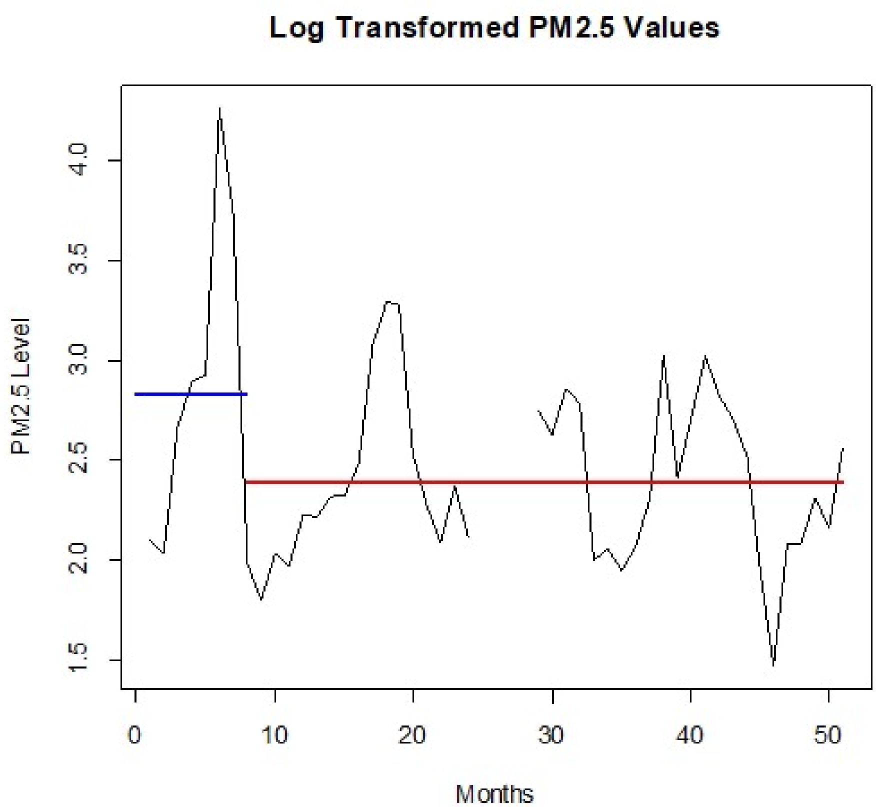

Figure 16: This graph overlays the average PM2.5 levels before and after the estimated change point on the log transformed monthly pollution data from Fresno.

if it was not significant, is the pollution regulation such as Rule 4901, implemented in 1993 and aimed at decreasing wood burning, that caused a decrease in PM10 emissions during the winter. Since the regulations were followed by a decrease in PM2.5 and PM10 levels, they are associated with an increase in air quality as less pollutants are emitted by human factors (Yap et al., 2015). One possibility for the lack of a statistically significant change may be the small sample size.

However, although Rule 4901 was implemented in 1993, it was not enforced until 2003 (Yap et al., 2015). Although decreases in air pollution were observed in 2000 and 2001, the lack of enforcement implies that another source of pollution may be the reason behind the decrease. In the San Joaquin Valley, levels of polycyclic aromatic hydrocarbons or PAH, a compound associated with agricultural burning, were found to have declined as well during this time period (Noth et al., 2021). This signifies that during 2000 and 2001, there were high levels of agricultural burning, and the subsequent reduction in later years may have been the reason behind lowered pollution rather than any effect caused by Rule 4901. 54% of annual agricultural burning occurred during the winter from January 2000-May 2019, strengthening the hypothesis that there may be a relation between these events.

The graphs below present information about the bootstrapped test statistics evaluated at each observation within the data set with the mean shift removed. These graphs were generated removing any potential mean shift utilizing the bootstrap method, meaning that the histograms should approximately follow the uniform distribution as stated by the null hypothesis. The graph for Fresno is presented first, followed by that for Silao. The probability of observing any specific value should be roughly equal to the others, as in the uniform distribution, under the null hypothesis. However the graphs indicate that this model has a higher probability of detecting change points at the beginning and end of the dataset, creating another potential problem when attempting to detect change points as the algorithm may lack the ability to correctly detect changes occurring in the middle or may falsely detect a nonexistent change at the beginning or end of the dataset. Thus, the type 1 error rate is inflated and there is a discrepancy between the nominal and actual rates of type 1 error.

7 Conclusion

In this project we developed a new method to analyze air particulate levels and successfully applied it to PM10 and PM2.5 data to determine if a mean shift in air pollution had occurred, basing the test statistic on the SST evaluated before and after each potential change point to determine significance. The bootstrap method was used to resample the residuals, which were modeled with ARMA(1,1), and add the estimated seasonality to approximate the model. This was necessary to estimate the residuals and to determine if the change seen could have been by chance as opposed to something unlikely to randomly occur. The test statistic was based on the SST evaluated at each step, The statistics greater than the value of the minimum change point were averaged, returning a p-value of 0 indicating that the shift is unlikely to have occurred by chance.

We found that there was not a statistically significant mean shift in Fresno for either metric, while in Silao there was a statistically significant shift in PM10 levels around 2015. These changes may be due to construction of factories and the expansion of the brickwork and automobile industries in Silao emitting fumes containing PM10 particulate matter, while in Fresno they may be linked

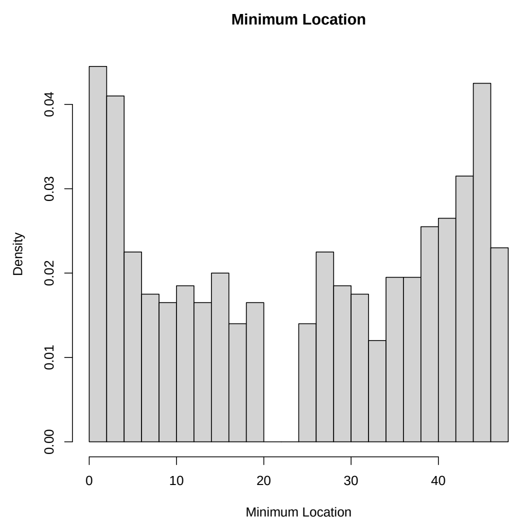

Figure 17: This graph presents the minimum location of the test statistics for the Fresno data.

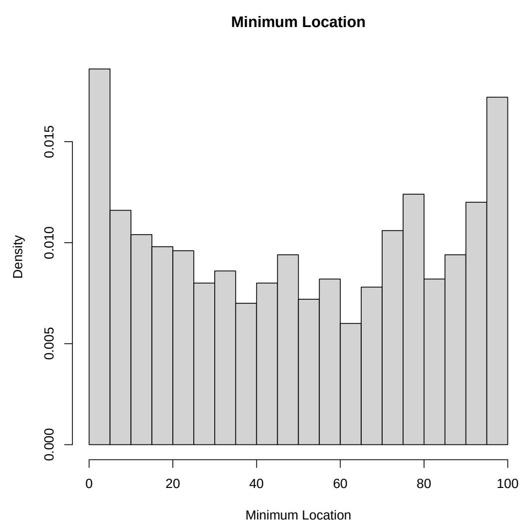

Figure 18: This graph presents the minimum location of the test statistics for the Silao data.

to an uptick in agricultural burning during 2000 and 2001, which was reduced in later years; other potential causes for the high PM2.5 levels may be related to emissions (de Foy et al., 2019). With regards to the PM10 levels within the San Joaquin Valley, agricultural activities were found to have a significant impact, such as fungal spores, plant detritus, and pollen grain, which were around 11-15% of PM10 mass in the fall of 2000 (Chow et al., 2015). As with the PM2.5 index, PM10 is also highly correlated with soil emissions, specifically the sand or clay content of the soil (Carvacho et al., 2001). Particulate nitrate, associated with soil emissions contributes heavily to such pollution in winter months (Ying et al., 2008). As soil emissions increase, predicted PM2.5 levels were seen to increase as well (Kleeman et al., 2019), and reducing oxides of nitrogen was found to be effective in controlling PM2.5 levels. Controlling oxides of sulfur is similarly effective but the lack of such compounds within the San Joaquin Valley signifies that efforts to reduce them have minimal effect (Chen et al., 2014). However, while reducing oxides of nitrogen can provide some decreases in PM2.5 levels, such gains are often offset by summer forest fires and residential wood burning in the winter, meaning that the cooler temperatures may provide an increase in PM2.5 that cannot be reduced by focusing on soil emissions. Fog is associated with the production of ammonium nitrate as well, which also contributes to PM2.5 emissions.

We hope to apply a similar methodology to other air pollution data to determine what other locations have experienced a mean shift in air pollution. Determining when such a shift may have occurred will be useful for identifying factors that would have contributed to the change and potentially preventing levels of pollutants from climbing higher or mitigating future differences in air pollution.

Acknowledgment

I would like to thank Dr. Kimihiro Noguchi for supervising me during this thesis and Dr. Olivia Chilian for suggesting the PM10 dataset in Silao and a relevant news article.

References

1. NASA/LARC/SD/ASDC. (2003). NARSTO EPA Supersite (SS) Fresno, tapered element oscillating microbalance (TEOM) Particulate Mass Concentration Data [Data set]. NASA Langley Atmospheric Science Data Center DAAC. Retrieved from https://doi.org/10.5067/ASDCDAAC/NARSTO/0046 on May 6, 2021.

2. Noguchi, K., Gel, Y.R., Duguay, C.R. (2011). Bootstrap-based tests for trends in hydrological time series, with application to ice phenology data. *Journal of Hydrology* 410, 150-161.

3. Chen, J., Gupta, A.K. (2011). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*, Second Edition. Springer, New York.

4. Goodwin, R., Noguchi, K. (2022). *Analysis of PM10 Levels in Silao, Mexico.* Poster Presentation, Scholars Week 2022, Bellingham, WA.

5. Poore, M.W. (2002). Levoglucosan in PM2.5 at the Fresno Supersite. *Journal of the Air & Waste Management Association*, 52:1, 3–4.

6. Noth, E.M. et al. (2020). Decrease in ambient polycyclic aromatic hydrocarbon concentrations in California's San Joaquin Valley 2000–2019. *Atmospheric Environment*, 242, 117818.

7. Secretaría de Medio Ambiente y Ordenamiento Territorial (SMAOT) PM10 Hourly Data. Retrieved from https://smaot.guanajuato.gob.mx/sitio/seica/historicos/silao on May 6, 2021.

8. Manufacturing In Guanajuato. Retrieved from https://napsintl.com/manufacturing-in-mexico/mexico-manufacturing-locations/manufacturing-in-guanajuato/ on May 9, 2022.

9. Silao es la sexta ciudad con más contaminación en México. Retrieved from
https://www.milenio.com/estados/silao-es-la-sexta-ciudad-con-mas-contaminacion-en-mexico
on May 6, 2021.

10. EPA PROPOSES TO RECLASSIFY CENTRAL VALLEY AIR FROM SERIOUS TO SEVERE. Retrieved from
https://www.epa.gov/archive/epapages/newsroom_archive/newsreleases/cbdb98a40b991a6a852570d8005e13d0.html
on 8 May 2023.

11. U.S EPA DOWNGRADES SAN JOAQUIN VALLEY AIR. Retrieved from
https://archive.epa.gov/epapages/newsroom_archive/newsreleases/ee00e5fd6dca7bbd852570d8005e145a.html
on 8 May 2023.

12. Modeling PM Nitrate Formation in the San Joaquin Valley Air Basin during Recent Years. Retrieved from
https://atmos.utah.edu/aquarius/presentations/55_kleeman.pdf
on 8 May 2023.

13. de Foy, Benjamin, Schauer, J.J. (2019). Changes in speciated PM2.5 concentrations in Fresno, California, due to NOx reductions and variations in diurnal emission profiles by day of week. *Elementa: Science of the Anthropocene*, 7: 45.

14. Chen, J., Lu, J., Avise, J.C., DaMassa, J.A., Kleeman, M.J., and Kaduwela, A.P. (2014). Seasonal modeling of PM2.5 in California's San Joaquin Valley. *Atmospheric Environment*, 92, 182–190.

15. Chow, Judith C. et al. Characterization of Ambient PM10 Bioaerosols in a California Agricultural Town. Retrieved from
https://aaqr.org/articles/aaqr-14-12-oa-0313.pdf
on 8 May 2023.

16. Carvacho, O.F. et al. RELATIONSHIP BETWEEN SAN JOAQUIN VALLEY SOIL TEXTURE AND PM10 EMISSION POTENTIAL USING THE UC

DAVIS DUST RESUSPENSION TEST CHAMBER. Retrieved from https://elibrary.asabe.org/abstract.asp?aid=7046 on 8 May 2023.

17. Meng, Y.-Y., Rull, R.P., Wilhelm, M., Lombardi, C., Balmes, J., Ritz, B. (2011). Outdoor air pollution and uncontrolled asthma in the San Joaquin Valley, California. *Journal of Epidemiology & Community Health*, 64, 142–147.

18. Yap, Poh-Sin et al. (2015). Effectiveness of Residential Wood-Burning Regulation on Decreasing Particulate Matter Levels and Hospitalizations in the San Joaquin Valley Air Basin *American Journal of Public Health*, Vol 105, No. 4

19. Qing, Y., Lu, J., Kaduwela, J., and Kleeman, M. (2008). Modeling air quality during the California Regional PM10/PM2.5 Air Quality Study (CPRAQS) using the UCD/CIT Source Oriented Air Quality Model – Part II. Regional source apportionment of primary airborne particulate matter. *Atmospheric Environment*, 42, 8967–8978.