MATH334 Group Project: California Pollution

A. Reid*, A.Nazarovs†, N. Hofmann‡ and R. Harwood§

*34747826
†34864008
‡34697217
§34734945

***Abstract*—Air pollutants cause harm to humans and the environment. Research suggests there is a level of pollution which when surpassed will cause more harmful effects. This report aims to use time series analysis and forecasting to determine if measures to reduce pollution are working. Focusing on nitrogen dioxide and ground level ozone pollution in LA, the Box-Jenkins approach is applied to the time series of the mean level of pollutants and the resulting models are used to forecast the future pollutants levels. Forecasts suggested that pollution was decreasing in LA. The results indicate that measures to reduce pollution from the state of California are working.**

I. Introduction

There are five common air pollutants: ground level ozone (O_3), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO) and particle matter [1]. The United States has been regularly monitoring the first four of these with stations situated across the entire country. This is a huge operation and thus takes sizeable time, effort and money, so why do it?

In other countries, studies have shown a link between air-borne pollutants and hospital admissions for conditions such as flu, pneumonia and cardiovascular disease conditions, which if left untreated could result in death [2]. One of the best ways to combat this issue would be to prevent exposure to pollutants, which could be achieved by reducing the amount present in our daily lives. This study in particular also mentions an interaction between NO_2 and O_3 which results in lower levels of O_3 being present in the winter. Upon further research, the exact reactions which take place are [3]:

$$\begin{aligned} NO_2 &\xrightarrow{\text{light}} NO + O \\ O + O_2 &\rightarrow O_3 \\ O_3 + NO &\rightarrow NO_2 + O_2 \end{aligned}$$

From this we can expect to see higher NO_2 levels during the winter; due to the reaction not having the prerequisite photon to take place. The set of reactions suggest an inverse relationship, one which could be interesting to look at given we ideally wish to reduce the levels of both pollutants.

The World Health Organisation (WHO) have guidelines set out to reduce pollution and have provided clear targets for ‘acceptable’ levels of pollution to be achieved. These targets include a mean of $100\mu g/m^3$ every 8 hours for O_3 and an annual mean of $40\mu g/m^3$ for NO_2 [4].

This report will aim to use time series analysis to assess the levels of NO_2 and O_3 and use forecasts to determine if they are reducing in line with the WHO air quality guidelines.

II. Methods

We initially had a relatively small subset of data, which contained the details of air pollution levels across the state of California in 2013, we searched this data for interesting ideas and patterns which we could analyse. Then we would see if they translated to a larger dataset for thorough analysis.

Upon research we found that California historically had a very high level of air pollution; smog was a common occurrence in Los Angeles (LA) across the 1940s and 50s. From the late 60s onwards there has been a state-wide effort to decrease these pollutant levels [5]. The effects of this should still be present in our data, so while we can locate yearly trends, the results should show the average pollution level decreasing across all the different pollutants. We note that this is something widely reported to be the case by the environmental protection agency (EPA) [6].

Initial Analysis

In the beginning, we modelled the mean gas levels over the year 2013 for a selection of cities, such as Fresno, searching for results of particular interest, however all results were as anticipated across the four seasons with O_3 levels being highest in the summer while the levels of NO_2 and CO were at their lowest; the levels of SO_2 varied across the year. We took this one step further and looked at the ACF and PACF of the data, but this also did not give results of much interest as they all followed the findings of our research.

It appeared that the trends between the NO_2 and CO levels were very similar throughout the year, with the O_3 level being the opposite. Having noted this trend, in both our data and research, this served as inspiration for examining the potential trends in the larger dataset.

To see if this was the case, and there was the expected correlation in our data, the first step of analysing the larger data was to create a correlation matrix using the mean values of the four types of gases across all areas of California in order to identify any strong results. The resulting matrix is as visible below in figure 1.

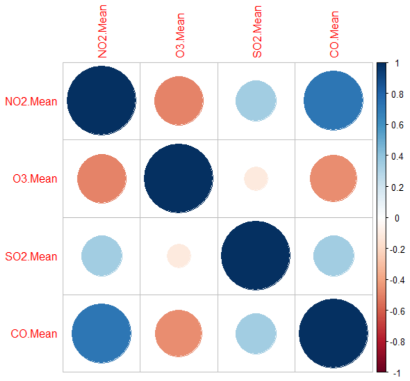

Fig. 1. A Correlation Matrix Across the Different Pollutants

This demonstrates that while there are relatively insignificant correlations between SO_2 and the other pollutants, there are strong correlations between the levels of NO_2 and CO as well as NO_2 and O_3. The former correlation was to be expected as their trends are very similar over a year. However, the strong negative correlation of the latter is a result of interest as this confirms our earlier suspicions of the two levels being dependant on one another. On the other hand, we should note that correlation and causation are not equal and so we cannot confirm that the equilibrium reaction mentioned before is a factor here.

The Dataset

The larger dataset contained the recorded daily pollutant levels for the four pollutants at sites located all across the USA; featuring data for varying lengths of time between 2000 and 2016. It became immediately clear that we needed to specify an area to analyse from the data and so we chose to use the city of LA. While the dataset supposedly had a large quantity of useful data, upon further inspection it was clear that the data was inconsistent and/or missing across all sites and years, so a careful selection of sites was required.

LA was interesting due to its historical air pollution and the data was better compared to most of the other sites in both depth and longevity. This led to the selection of sites 1103 and 5005, one of which is located near an airport and the other in the city. This means they would experience different pollutant levels which creates an interesting variable for comparison. Figure 2 shows the mean trends over the given time period.

With site 1103 in blue and site 5005 in red, with their rolling average being shown in cyan and pink respectively, these plots clearly demonstrate the inverse relationship between NO_2 and O_3. Another feature is that the levels collected at site 5005 are consistently different of site 1103.

This, therefore, gives us a reason to further analyse this data as there are multiple factors to explore and test for trends as well as providing a sufficient background of historic data from which we can produce forecasts to use for comparison against previous theories and hypotheses.

However, we can observe several gaps in the data. To minimise the number of gaps, the period from 2004-04-25 to 2010-11-20 was chosen. The forecast testing set was chosen to be from 2010-11-21 to 2011-11-19.

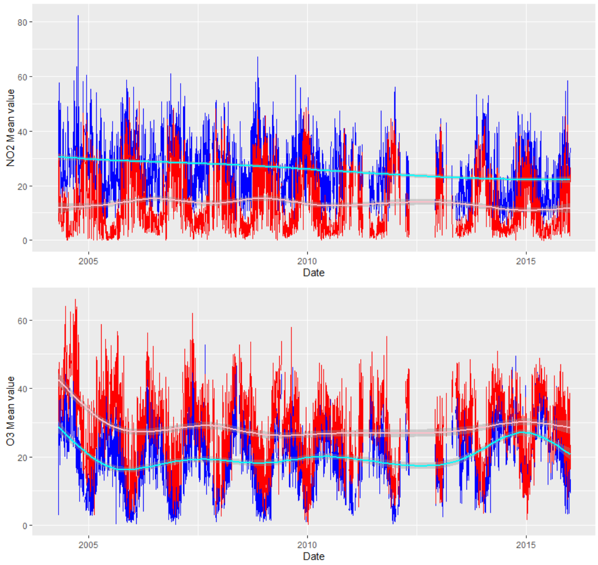

Fig. 2. Two Plots Showing the Average Pollution Level

Mathematical Methodology

We have already noticed the yearly seasonality of the NO_2 and O_3 level. The ACF confirms this by its wave structure with approximately 1-year period. However, if we look closer, smaller “waves” with 1-week period can be identified around the yearly ones. This is particularly apparent for the NO_2 level at site 1103 (as can be seen within figure 3) but is present across all our models.

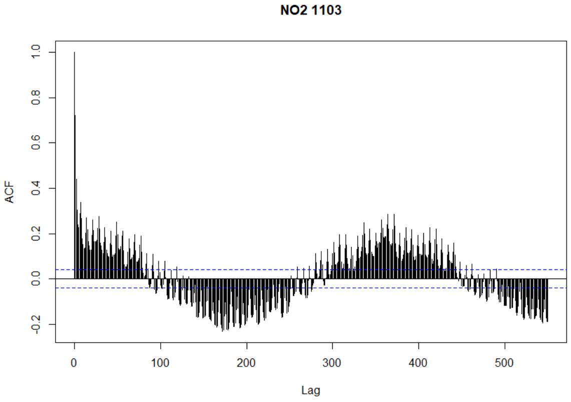

Fig. 3. ACF Plot of the NO_2 Level at Site 1103

This ‘double’, or complex, seasonality is an issue as the usual SARIMA model does not account for this. Also, the `arima()` function in `R` does not allow for periods larger than 350. However, there are several ways of dealing with it. For instance, we could use STL (Seasonal and Trend decomposition using LOESS (locally estimated scatterplot smoothing) which is extended to treat multiple seasonality by the `mstl()` function in `R`. Another way is to use Fourier series [7]. Even though these methods can be quite easily applied in `R`, their mathematical interpretation is outside our expertise, so we will follow our own method.

We begin by simulating the missing data. This could be done by averaging the neighbouring data points. However, the more mathematically correct way is to use Kalman Smoothing or 'imputing' (`na_kalman` function within the `imputeTS` package in `R`). We then calculate the weekly mean values for each pollutant. This can be done either by manipulating the data frame with the `dplyr` package or with a simple `for` loop. Such averaging helps in several ways. First of all, we get rid of complex seasonality so that only the yearly seasonality is present. Secondly, the period is reduced from 365 to 52, which allows us to use the `arima()` function in `R`. Thirdly, the imputed missing data (which approximates the underlying model) affects the model choosing less.

This method has its drawbacks. Clearly, the exact period is a bit bigger than 52. The effect of this is not major in our case as our horizon is only around 6.5 years, but creates another disadvantage as we now have 343 data points, which is not much for time series analysis because it is based on the infinite past assumption. On the one hand, this reduces random variability in data, but at the same time neglects potentially important exogenous shocks (e.g. public events, extreme weather conditions, etc.) that could help estimate our model better.

Beyond that, it may seem like our predictions will only be about the weekly mean level instead of the daily level, but this can be overcome by using the average levels of days of the week. This means that if we empirically deduce their relative distribution in the long-run, we can expect a given week to follow the same pattern in the absence of previously known exogenous shocks. In the end, we can just take our weekly forecast and split it into 7 days, according to this distribution. The following plot, figure 4, shows the average level for each day of the week, emphasising the difference between weekdays and weekends, the main cause of weekly seasonality.

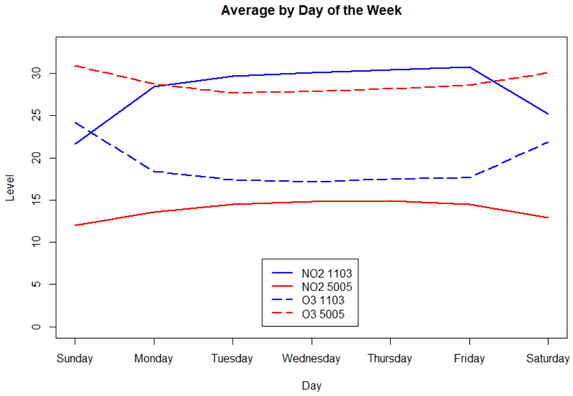

Fig. 4. The Average Pollutant Level by Day of the Week

We start our model-building by differencing our weekly data with lag 52 and run an ADF test to check if our differenced data is stationary. Once the data is stationary, we can then look at the ACF and PACF plots and fit the appropriate models as required.

To fit these models, we use the appropriate lag and differencing values as well as taking a range of apparently applicable values from the ACF and PACF, before iteratively checking each of these according to Box-Jenkins Approach.

We note the Box-Jenkins approach is an iterative way of finding the most suitable model to fit to our data by assessing which models appear to be appropriate by reading the ACF and PACF to obtain a selection of potential parameter sets and performing a Ljung-Box test on each to find their p-values before taking those with the result closest to 1. The Ljung-Box test is a way of assessing residuals for normality. We also check how good our models by comparing their forecasting potency, using such metrics as MAE and RMSE [8]. We can then choose the models with the lowest AIC scores, highest p-values and smallest forecasting errors to be our models.

Model Fitting

As mentioned before, the data was differenced once with lag 52 and then the differenced data was tested for stationarity through the ADF test. This test showed strong evidence in favour of the alternative hypothesis (stationarity), so we did not difference our data anymore.

We then considered the ACF and PACF of the differenced data. For both pollutants at both sites, the ACF showed peaks at lag 52 and appeared to be cutting off (i.e. having non-significant values at multiples of lag 52) while the PACF seemed to decay gradually over seasonal lags. This was a clear sign of the seasonal MA order being 1 and the seasonal AR order being 0, i.e. (0,1,1) seasonal component. Although, the ACF of O_3 at site 1103 is also noticeable at seasonal lags (could be decaying instead of cutting-off), which could be a weak sign of (1,1,1) seasonal component, so we checked models with that too.

Regarding non-seasonal orders, neither of the ACF nor PACF showed significant values at first few lags for the NO_2 levels at station 1103, so we will try only models of orders of 0 or 1. At station 5005 though, PACF of the NO_2 has a peak at lag 1 and then cuts off while ACF decays after its peak at lag 1. This is an obvious sign of AR(1) model. For O_3, everything gets much more complicated as its ACF and PACF show considerable peaks up to lag 5 and even up to lag 7 in one case (ACF of O_3 series at site 5005).

Therefore, for the NO_2 levels, we have only 3-4 models to consider. However, to enforce the Box-Jenkins approach for O_3 levels, as there are potentially a lot of models to check, we create a matrix with rows and columns representing AR order (p) and MA order (q) respectively, while each entry representing AIC of model with corresponding orders fitted. An AIC matrix example is given below:

	MA=0	MA=1	MA=2	MA=3	MA=4	MA=5
AR=0	-2320.748	-2350.851	-2367.295	-2365.706	-2364.071	-2365.614
AR=1	-2364.550	-2373.273	-2375.469	-2376.438	-2377.151	-2378.189
AR=2	-2369.534	-2367.756	-2369.594	-2375.475	-2377.280	-2376.205
AR=3	-2367.551	-2365.860	-2373.049	-2377.041	-2376.203	-2374.242
AR=4	-2370.119	-2378.765	-2377.892	-2376.099	-2373.420	-2373.954
AR=5	-2375.646	-2378.004	-2376.200	-2374.570	-2372.640	-2373.963
AR=6	-2376.132	-2376.004	-2374.454	-2375.072	-2379.760	-2371.881

III. Results

Once the data was converted into a time series of weekly values for the years 2004 to 2010 (getting rid of double seasonality), the model-fitting process could take place. We could then compare the pollutant levels at the two different sites with the outlined recommended WHO guidelines. These time series are as follows in figure 5.

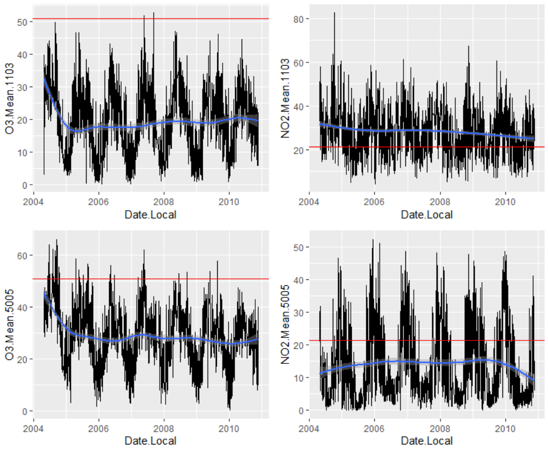

Fig. 5. The Four Time Series of LA Pollutants

By adding in the rolling average trend lines in blue and the WHO recommended guideline in red, we can see that the rolling yearly averages are decreasing towards the WHO standards or are already below, meaning that the respective weekly average pollutant levels are going down and the standards will be met. This meets our expectations that the pollutant levels will be decreasing, but we can also see that the negative correlation is apparent throughout each area; even with the rolling averages decreasing. An example of this is at the beginning of each year, we see that the O_3 levels are around their lowest point and the NO_2 levels are around their highest, which when we consider this to be in the winter, makes logical sense. Another apparent result to note is that the O_3 levels appear to be drastically lower than the WHO guideline, in comparison to those of NO_2, across both sites.

The resulting models from fitting were as follows. For the NO_2 levels, we found the models to be SARIMA(1,0,0, 0,1,1, 52) for both the 1103 and 5005 sites in LA, their AIC values being 1941.938 and 1799.699 respectively. We found the most appropriate models to be SARIMA(6,0,4, 0,1,1, 52) for O_3 at site 1103 (AIC = -2379.76) and SARIMA(6,0,0, 0,1,1, 52) for O_3 at site 5005 (AIC = -2165.118).

Using these models, we have been able to generate forecasts with the lowest possible errors for the pollution levels at each site. These can be seen in figure 6.

The forecasting line is in dark blue, with 80% and 95% confidence intervals being apparent in lighter shades. These show that we have the same trend being predicted into 2011 for the pollutant levels, with the average still appearing to be decreasing or consistenly low enough. We can gauge this by the decrease of the peaks in each of these forecasts; suggesting a further decrease of pollutant levels in 2011.

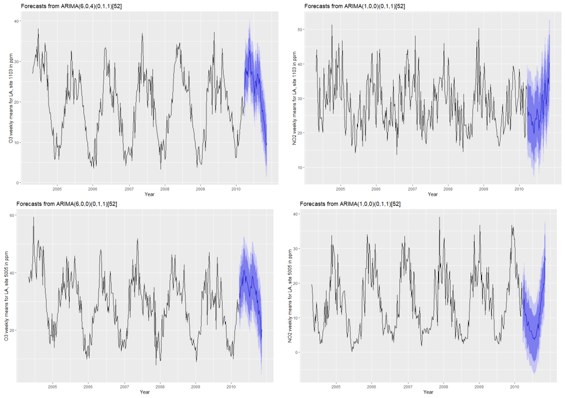

Fig. 6. The Four Forecasts for LA Pollutants

Upon comparison of these forecasts with the data we have available, we are able to find the mean absolute error, which is the mean difference between the forecasts and the actual observations [8]. These values for the O_3 and NO_2 levels respectively for site 1103 are 3.062 and 5.036 and for site 5005 are 5.090 and 4.572; with each rounded to three decimal places. We also find that the root mean square errors are 3.672, 6.446, 6.034 and 4.572 analogously. These values compared to those of alternative models show that our models produce a good forecast for the 2011 period as each of these values are relatively low, implying a low error of the predictions.

IV. Discussion

Our results suggest that for NO_2 and O_3 at both sites the mean level of pollution is decreasing or consistently below the guidelines and should it continue in this manner the WHO guidelines will be met or continue to be upheld. We also notice that the O_3 level at site 5005 is the only one above the guidelines and will probably take a few years to hit the necessary value.

The accuracy of our results is limited by the data we received, there were several rather large gaps in the data which forced us to use a fraction of what was available. We also needed to impute data and average weeks which reduced the purity of our data, however, did not compromise it too much. If we needed a daily forecast, we could modify our code to split weekly forecasts according to the relative distribution of average daily levels over the week.

With more time and a complete dataset spanning over a longer period of time we could produce more accurate forecasts spanning over a longer time which would help us notice whether site 5005's O_3 level would in fact drop low enough to be considered safe. We could even make a forecast for up to 2021.

References

[1] "Common air pollutants and their health effects," 04 2013. [Online]. Available: https://www.health.nsw.gov.au/environment/air/Pages/common-air-pollutants.aspx

[2] T. W. Wong, T. S. Lau, T. S. Yu, A. Neller, S. L. Wong, W. Tam, and S. W. Pang, "Air pollution and hospital admissions for respiratory and cardiovascular diseases in hong kong." *Occupational and Environmental Medicine*, vol. 56, no. 10, pp. 679–683, 1999. [Online]. Available: https://oem.bmj.com/content/56/10/679

[3] K. Kenty, N. Poor, K. Kronmiller, W. McClenny, C. King, T. Atkeson, and S. Campbell, "Application of CALINE4 to roadside NO/NO2 transformations," *Atmospheric Environment*, vol. 41, pp. 4270–4280, 06 2007.

[4] "WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulphur dioxide," 2005. [Online]. Available: https://www.who.int/airpollution/publications/aqg2005/en/

[5] "History," 2021. [Online]. Available: https://ww2.arb.ca.gov/about/history

[6] "Air quality - national summary," 11 2020. [Online]. Available: https://www.health.nsw.gov.au/environment/air/Pages/common-air-pollutants.aspx

[7] Hyndman, R.J., and Athanasopoulos, G., ***'Complex seasonality'*** *in Forecasting: Principles and Practice*, 3rd ed. OTexts: Melbourne, Australia, 2021, accessed on 05.02.2021. [Online]. Available: https://otexts.com/fpp3/complexseasonality.html

[8] A. Gibberd, "Time Series (MATH 334)," 2021, Chapter 6.4.