Forecasting NO_2 in Areas of Varying Population in the State of California

G.Pollard*, S.Holdstock†, A.Blakeley‡ and Y.Li§
*34789359 †37526685 ‡34605533 §34740473

***Abstract*—Epidemiological studies provide some evidence that long-term NO_2 exposure may decrease lung function and increase the risk of respiratory symptoms, so the fluctuation of this pollutant in built-up areas has been of growing concern to many health organisations across the US in the past decade[1]. Motivated by these findings, we conducted a time series analysis utilising data of the concentration of NO_2 in the state of California from 2008 to 2016 for two specific geographical areas, the city of LA and the countryside. Justifications were made for a SARIMA$(5,1,4,0,0,1,12)$ model to describe the LA data, and another model was generated using harmonic regression for the data regarding more rural communities. It was shown that NO_2 levels were drastically lower in the countryside compared to LA. We also found that NO_2 pollution was a lot more variable in LA than in the countryside, spiking so much in fact that plotting monthly data was more feasible to reduce noise and variance in our data points. Both datasets exhibited annual seasonality, peaking in winter and falling in summer, why this occurs has been studied to a great extent in other report papers[2].**

I. Introduction

A. What is Air Pollution?

Air Pollution is the release of various gases into the atmosphere. These gases can exceed the natural threshold which the environment dilutes, dissipates, or absorbs them. The gradual increase of these pollutants can negatively affect the environment by reaching concentrations that can cause unwanted consequences, including damages to our health and economy. There are six major air pollutants that are used as indicators of air quality. These include carbon monoxide, nitrogen oxides, sulphur dioxide, ozone, particulate matter, and lead. The three air pollutants that are particularly important (especially in urban settings) include sulphur dioxide, nitrogen dioxide, and carbon monoxide. These gaseous chemicals are released directly into the atmosphere from the combustion of fossil fuels such as petrol and natural gas. Also, ozone is another gaseous contaminant formed within the atmosphere from the chemical reaction between nitrogen dioxide and a variety of volatile compounds including petrol vapours[3].

B. The Difference Between Urban and Rural Pollution

An urban area is a location that has a high population density. These typically include artificial cities which can explain their extensive infrastructure. Whilst a rural area is an environment that is mostly natural and located away from crowded regions. We can deduce that urban areas experience worse rates of air pollution compared to rural areas. One reason is because urban areas are more populated. To take a few examples, this means that there are more vehicles on the road (which is the largest factor of poor air quality); higher rates of energy production; increased factory emissions; and so on. All these components produce the chemicals that cause an over-polluted atmosphere[4].

C. Seasonality of Air Pollution

When air pollution is analysed by seasonality, there are higher levels of pollution during the winter as compared to the relatively lower levels in the summer months. The higher levels of pollution in the winter months are linked with household heating appliances releasing more emissions due to the colder weather. Another contributing factor is the increase in road vehicles being used to avoid the lower temperatures. We tend to record higher levels of air pollution in the winter due to natural phenomena, such as overcast weather. That is, a kind of temperature inversion where the warmer air above the clouds act as a barrier from air pollutants[5].
The exception to all of this is ozone. The creation of ozone is at its highest levels of concentration when there are longer days and higher temperatures. And the reduced levels of nitrogen dioxide and carbon monoxide in the summer months contribute to photochemical reactions occurring due to solar radiation, which causes the formation of ozone[6].

D. Use of Time Series Analysis

Time Series Analysis is the process of analysing a collection of data points organised in time. It is generally used to clean, understand, and forecast given data[7].
Using this type of analysis works well with the data set that we analysed, as we have a large amount of daily readings spanning a long period of time. Models used in time series analysis can be used in understanding the true meaning of the given data set, so we can predict when data points can be expected[8].

II. Methods

A. Preliminary Analysis

1) Describing the Data: A major trend to investigate regarding air pollution seemed to be to analyse how the density of pollutants changes as time passes in areas of high population (such as in a city), and areas of low population (like those in more rural communities). Due to the scope of this report, we focused on two specific datasets; one detailing the average NO_2 concentration around Los Angeles for each day in the years 2008-2016, and the other referring to the same measurements but taken from less densely populated rural areas in the same years. The data collected in Los Angeles will be referred to as the "LA data" whereas the more rural data will be referred to as "NIAC data" (meaning not in a city). We only made our analysis on data from 2008-2015 so that we could use the 2016 data for testing later on. The aim of the study was to construct a model that could predict

Group Report for MATH334 Time-Series Analysis 2020/21, Group: PRES01/01/1

NO_2 concentrations for future days, months and years.

2) Cleaning and Plotting the Data: Every 1 in 4 records had N/A values (not applicable), and some records over long periods simply didn't exist, and so averages were taken to remedy this. Figure 1 shows a plot of the corrected, raw data.

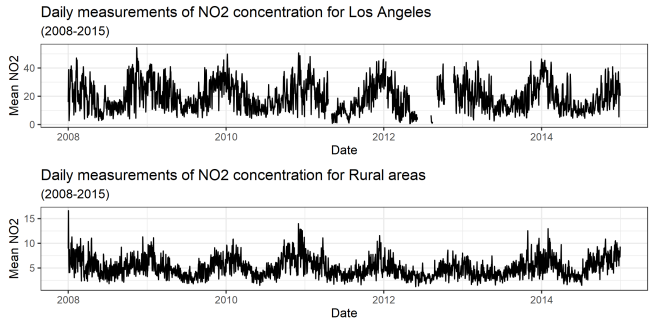

Fig. 1. A time series of both datasets, please note that breaks in the series lines were intentional, and are a result of missing data.

We saw that for both datasets there appeared to be some annual seasonality since pollution dipped around summer and raised during winter, this lined up with our pre-existing research relating to similar trends in other cities[9]. Subsequently, we were interested in modelling yearly patterns for our data, however we knew from our research into external information sources surrounding this task[10], that modelling long term trends with daily data can be tricky. To solve this problem, we 'compressed' our data into months by taking averages, this also helped us approximate missing values such as those in 2012. Although we lost the daily patterns, we sharply reduced variance of our data and averaged out noise; this allowed us to make more accurate forecasts that take into account seasonality. Notice that in Figure 2 the periodicity remains.

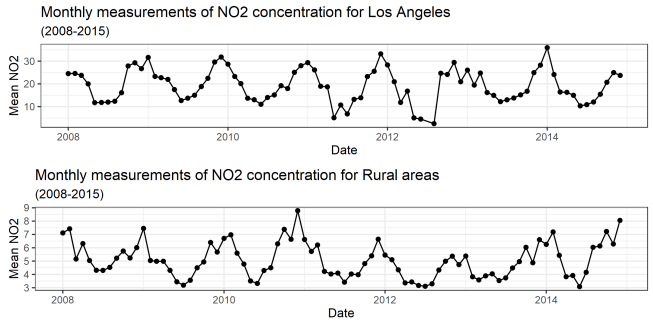

Fig. 2. A new time series of monthly data.

B. Model Fitting

1) Fitting Using A (S)ARIMA Model: The data we had chosen appears cyclostationary, as the mean was not independent of time; for example, the expected value of both series in winter was higher than in summer by about 50%. However, the variance for both datasets changes over the course of each year and so simply claiming this stationarity is an error. A quick inspection of the ACFs indicated that differencing was required due a very slow decay. We differenced our data by 1 unit of time, and an ADF test was used to prove a rejection of the null hypothesis that our differenced datasets had a unit root, and so this further supported stationarity. Figure 3 shows the transformed time series.

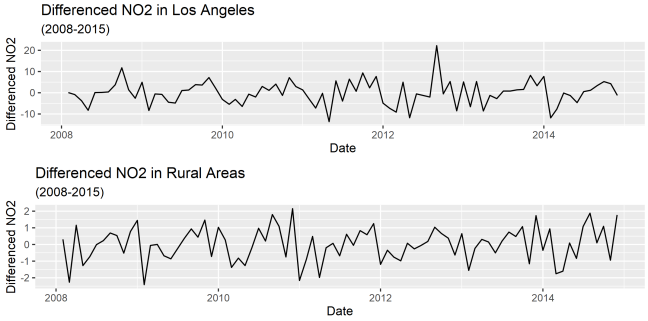

Fig. 3. A time series of the differenced data.

After this transformation, the time series produced were stationary and had a fairly constant variance. Because we had differenced the data, a suitable model for forecasting was an ARIMA or SARIMA model, and so the next task was deciding between these options and then finding the optimal model structure for both datasets. The ACF and PACF for our transformed data shed light on whether we required a seasonal model. Figure 4 shows the ACF and PACF for both transfromed, stationary datasets. The ACFs and PACFs both

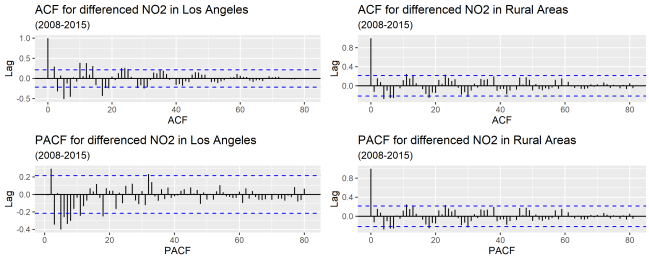

Fig. 4. The ACF and PACF for both datasets after transformation.

decay rather quickly, which solidified our hypothesis that an ARIMA model would fit our data well. A simple AR(p) and MA(q) was rejected for our rural data because both the ACF and PACF had a strong geometric decay. It was also noted that even though our data was differenced, it still appeared to have some seasonality, as lags at multiples of 12 were occasionally significant, this could suggest an annual trend. Therefore, to keep our p and q low but retain the yearly seasonal relations, we decided that a SARIMA model could fit our data and should be considered too.

2) Finding the Structure of the Model: We then had the task of finding the parameters for these proposed models. To overcome this problem, we iteratively tested different models using the Box-Jenkins approach. In order to evaluate how our models performed in these tests, we looked for the following features:

- Standardised residuals that look like white noise.
- Fast decaying ACF of the residuals.
- High p-values for the Ljung-Box test statistic at large lags.
- Low AIC value.
- High log-likelihood.
- Small model size (p+q)

We tested all ARIMA$(p, 1, q)$ models up to $p, q = 8$ for both time series. For the LA data, the most optimal model was ARIMA$(5, 1, 4)$ as it had the lowest AIC and nearly the highest log-likelihood while also being relatively simple. In addition, the standardised residuals appear to be white noise, and the Ljung-Box statistic is above 0.4 for all lags. For the NIAC data, the highest log-likelihood belonged to the most complex model ARIMA$(8, 1, 8)$; however, most of the simpler models showed serial correlation between residuals in the Ljung-Box test, and so a compromise through the model ARIMA$(4, 1, 2)$ was made. This model performed very well in most aspects of the list detailed above, including having the lowest AIC, and so it was used to describe the NIAC dataset. To find our SARIMA model, a seasonal component was added to the LA model, this increased the log-likelihood and Ljung-Box statistic whilst decreasing the AIC. This made sense because the data originally appeared to have a strong seasonality; so we decided to keep this SARIMA component for the LA data. The NIAC model was already performing very well and, subsequently, a seasonal component did not improve the log-likelihood enough for us to justify keeping it. Figure 6 showcases the parameters of both models and other relevant measurements.

Area	Model	$\hat{\sigma}^2$	Log-Likelihood	AIC
Los Angeles	ARIMA$(5, 1, 4)$	9.151	-213.69	447.37
Los Angeles	SARIMA$(5, 1, 4)(0, 0, 1)[12]$	8.649	-212.17	446.35
Rural	ARIMA$(4, 1, 2)$	0.518	-94.94	203.87

Model	ar1	ar2	ar3	ar4	ar5	ma1	ma2	ma3	ma4	sma1
SARIMA$(5, 1, 4)(1, 0, 0)[12]$	-0.37	0.96	0.29	-0.81	-0.49	-0.22	-1.06	0.38	0.66	-0.20
ARIMA$(4, 1, 2)$	1.10	-0.18	-0.18	-0.26	NA	-1.76	1.00	NA	NA	NA

Fig. 5. Tables of the output from the arima() function containing relevant info regarding all selected models.

3) Fitting Using Harmonic Regression: We also considered a harmonic regression approach. To do this we use Fourier series. A Fourier series is a periodic function consisting of a summation of weighted sinusoids which are harmonically related (i.e. their frequencies are all multiples of the fundamental frequency). Fourier series are often used to approximate functions. This approximation can be extended and used to forecast future values for a time series.

For this to work properly, the time series we are modelling should be periodic. Our monthly NO_2 measurements in Figure 2 appeared to be very sinusoidal - peaking in winter and dipping in summer. Subsequently, we can find a Fourier series that approximates our training data well

We applied this technique to our data and removed any trigonometric coefficients which were not significant. This left us with the equations below, where p_t represents our harmonic regression for Los Angeles data and q_t represents our harmonics regression for rural data. See Eq 1.

$$\begin{aligned} p_t &= 20.7 + 8.9\cos(\frac{\pi}{6}t) + 1.4\sin(\frac{\pi}{6}t) - 0.04t \\ q_t &= 5.07 + 1.4\cos(\frac{\pi}{6}t) \end{aligned} \tag{1}$$

These forecasts have the added benefit of being incredibly simple and can be easily interpreted. Examples of these interpretations were:

- Our LA forecast was decreasing over time, whereas our rural forecast was not.
- Our LA forecast has a higher mean.
- Our LA forecast has a higher variance.

4) Fitting Using General Models: It would also be useful to generalise our model to other areas of the US. If the same model worked for both our urban and rural data, it could be a sign that it might have been effective at modelling data for any region. Unfortunately, our harmonic regression was not generalisable. However, it's possible that our ARIMA models were. We took both our ARIMA and SARIMA model and fitted the other dataset to each model. The coefficients shown in Figure 6 appeared to be strongly correlated with each other. This was a benefit of differencing our model, as it centred our data about 0 and subsequently allowed for the same model to fit different data well. This made sense because both datasets shared many features and relations from one month to the next. Further analysis here is required to test this hypothesis.

Data	Model	ar1	ar2	ar3	ar4	ar5	ma1	ma2	ma3	ma4	sma1
Los Angeles	ARIMA$(4, 1, 2)$	1.10	0.02	-0.43	-0.12	NA	-1.88	1.00	NA	NA	NA
Rural	ARIMA$(4, 1, 2)$	1.10	-0.18	-0.18	-0.26	NA	-1.76	1.00	NA	NA	NA
Los Angeles	SARIMA$(5, 1, 4)(1, 0, 0)[12]$	-0.37	0.96	0.29	-0.81	-0.49	-0.22	-1.06	0.38	0.66	-0.20
Rural	SARIMA$(5, 1, 4)(1, 0, 0)[12]$	-0.24	0.52	0.54	-0.65	-0.56	-0.25	-0.56	-0.72	0.66	-0.15

Fig. 6. Tables of the output from the arima() function containing relevant info regarding all models. Notice the correlation between all coefficients indicating a general model could work.

III. Results

A. Model Evaluation

The main avenue to evaluate our models was to train them on the first 85% of the data, and test on the latter 15%. We plotted our forecasted points against the test data which had been kept separate until then. Figure 7 shows how the models predict future data points. The shaded region represents our 95% confidence interval in which the model predicts future points will lie between.

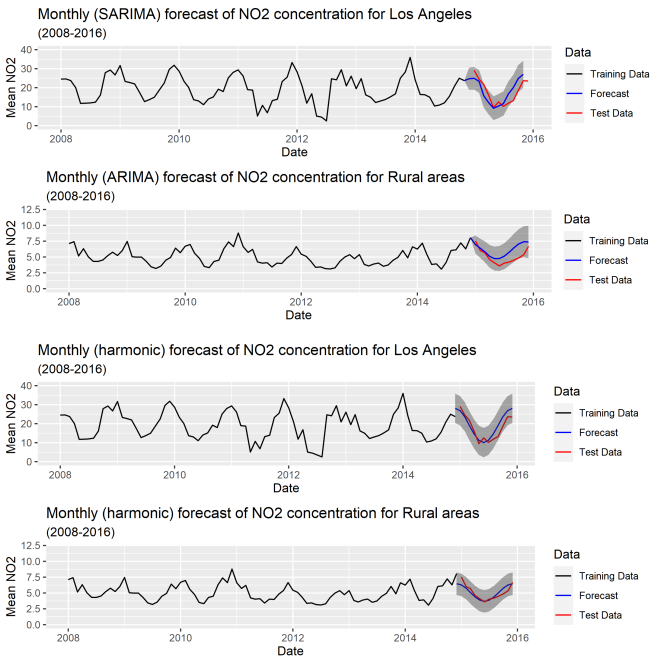

Fig. 7. Forecasts generated by models for the Los Angeles and Rural data respectively.

The models did an impressive job at forecasting future data points. All predictions line up closely with the test data, which was a sign that our models have successfully understood key

elements of the training data. Our models had encapsulated the seasonality of the data, as there were clear peaks in winter and dips in the summer. Useful metrics to compare the performance of all models were calculated; the table in Figure 8 describes them. These error metrics indicated that the Los Angeles SARIMA model and both harmonic models did an impressive job at forecasting future NO_2 monthly averages. However, the rural ARIMA model performed the poorly in comparison.

	Los Angeles (LA)		Rural (NIAC)	
Metric	SARIMA	Harmonic	ARIMA	Harmonic
MAE	1.889363	1.680275	1.082381	0.4736822
RMSE	2.379985	2.340504	1.277024	0.5900124
MAPE	0.1150727	0.1353198	0.2316965	0.08984376

Fig. 8. A table of mean absolute error, root mean square error, and mean absolute percentage error for all models.

We selected our final models based off their performance for the MAPE metric, as it compares the errors relative to the size of the estimates. Therefore, we choose the SARIMA model for the Los Angeles data and the harmonic model for the rural data. Final long term forecasts are shown in Figure 9.

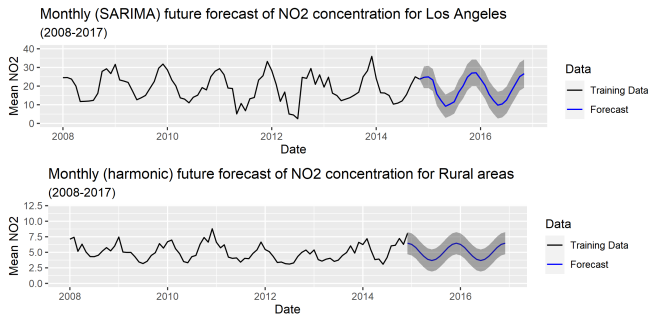

Fig. 9. Long term forecasts of our selected models

IV. Discussion

A. What do the Results Mean?

By simply observing the axes in Figure 1, it's clear that pollution is much greater outside of built up areas such as Los Angeles than in rural areas. Upon constructing appropriate models we found that there is strong periodicity in our data which matched from our literature review[11]. Our LA confidence intervals were rather narrow and its MAPE was low; this means that we can make accurate forecasts into the future, factoring in seasonality, to estimate what NO_2 concentrations are likely to be for a given month. In addition, we can also come to the conclusion that areas of population with a higher density have more variable pollution statistics, as we've seen in our analysis, this has made modelling tricky at times and decreases confidence in our predictions for more urban areas.

B. Limitations of the Analysis

1) Missing Values: The initial dataset consisted of 4 readings per day, in which only 1 had no missing values (notated as N/A). Furthermore, both datasets included some days where not a single reading was made, this means our analysis was not perfectly reflective of the true patterns these data exhibit. It could also mean that the method used to collect this data is unreliable.

2) Negative Values: Some of our values for the NIAC dataset were negative, this should be impossible as it doesn't make sense to measure a negative quantity of pollutant molecules in the air. Although there were not many of these values, and they were promptly removed and dealt with, this could once again suggest our analysis is untrustworthy to some extent.

3) Smaller Periodic Trends: Evidence from other studies [11] indicates that NO_2 fluctuates throughout the day, peaking twice at around rush hour. The Nyquist theorem tells us that to accurately reconstruct a signal, our data should be sampled at twice the rate of the period. Because our data was sampled daily, we missed this pattern and cannot further investigate it.

4) External Variables: The concentrations of NO_2 measured in the atmosphere may be heavily affected by other variables which have not been included in our data, such as wind speeds and temperature.

C. What Could we do Differently?

1) Other Approaches to Analysis: We could have delved further into analyzing how strong the periodicity of our data was, instead of just assuming by inspection that it was very periodic. For example, an investigation into any cyclic patterns that may be present using spectra and periodograms could have been very useful.

2) Checking for Dependency in Residuals: It may have also been worth investigating the dependence of our series of error terms for both models and checking for serial autocorrelation in more detail. For example, simply transforming the white noise series to emphasise peaks and troughs may have produced evidence for an ARCH model, this idea is also supported by the fact that our pollutant data is non-zero, and so it would suit this approach well.

3) Correlated Variables: It was apparent from the data and our literature review that levels of some pollutants can affect each other. For example, it would have been interesting to investigate in more detail the inverse relation between NO_2 and O_3 - would it be possible to predict NO_2 when given the O_3 levels?

REFERENCES

[1] R. Atkinson, "Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide," 2003.

[2] S. at AccuWeather, "Why air pollution is worse in winter," 2020.

[3] J. A. Nathanson, "Air pollution," 2020.

[4] M. Burdett, "Air pollution," 2019.

[5] T. at Airlief, "Why is air pollution worse during winter," 2019.

[6] W. F. Robert Cichowicz, Grzegorz Wielgosinski, "Dispersion of atmospheric air pollution in summer and winter season," 2017.

[7] T. Bush, "Time-series analysis: Definition, benefits, models," 2020.

[8] M. K. Douglas Montgomery, Cheryl Jennings, *Introduction to time series analysis and forecasting (Second ed., Wiley series in probability and statistics)*, 2015.

[9] Y. G. Dawn Roberts-Semple, Fei Song, "Seasonal characteristics of ambient nitrogen oxides and ground–level ozone in metropolitan north-eastern new jersey," 2012.

[10] B. Crocker, "How to predict a time series part 1," 2019.

[11] M. D. THOMAS and GILBERT, "Nitric oxide and nitrogen dioxide concentrations near the ground at menlo park, california," 2012.