ATENEO DE MANILA UNIVERSITY

PREDICTING TEMPERATURE FROM AIR QUALITY DATA IN ITALY

TECHNICAL PAPER

Course Code MATH 62.2

COURSE INSTRUCTOR **Mr. Ramil T. Bataller**

SUBMITTED ON May 30, 2022

BY Group 3

Carlo Gabriel M. Pastor
David Demitri Africa
Mark Kevin A. Ong Yiu

FOLLOWING THE REQUIREMENTS SET DURING THE ACADEMIC TERM 2021-2022 S2

Introduction

The following paper presents a statistical analysis of the *Air Quality Dataset* taken from the University of California Irvine Machine Learning Repository, which contains road-level pollution and weather data from Italy. Motivated by the extensively documented phenomenon of global warming and its relationship with air pollution, the goal of the paper was to build a multiple linear regression model with ARMA error terms that modeled daily temperature using the different types of emissions as predictors. This was accomplished in R through the following process.

First, various multiple linear regression models were compared against each other to determine an optimal base regression model, from which initial insights were also gathered. Second, time series analysis was performed over the residuals of the base model in order to determine an appropriate ARMA model for said residuals. The ARMA model and the base model were then combined and subsequently compared against the standalone base model to determine whether the model performance was improved by the inclusion of ARMA error terms.

It was discovered that over the span of one year, temperature counterintuitively decreased with the increased presence of air pollutants; however, much of the variance remained uncaptured until the ARMA error terms were included, suggesting that much of the relationship is governed strongly by time-dependent trends. It is suggested that pollution-generating activities like driving and indoor heating, as well as the outlier event of Mt. Etna's eruption right before the autumn and winter seasons, would significantly skew the amount of air pollutants during colder periods.

Overall, the authors recommend that further research be done on models that accommodate seasonality and on pollution datasets that span for longer time-periods, as average daily temperature is known to change with the seasons (i.e., higher temperatures in summer and lower temperatures in winter).

Background

Air quality refers to the measure of cleanliness and pollution in the air,[1] and it is a major target of concern for several Sustainable Development Goals (SDG).[2] Air that is contaminated by pollutants and other agents that modify the natural characteristics of the atmosphere can seriously affect long-term health outcomes through respiratory diseases and other illnesses.[3] The World Health Organization (WHO) estimates that almost 99% of the global population breathe air that exceeds WHO guideline limits on pollutants,[4] and that pollution in total leads to 7 million premature deaths every year.[5]

Relevantly, air pollution and global warming have had a long and well-documented relationship. Many studies have identified that pollutants such as methane (CH_4) and black carbon contribute heavily to the greenhouse effect — leading to the rise of temperatures.[6] Complex chemical and physical interactions due to the introduction of additional contaminants can also result in changes to the energy balance in the atmosphere, resulting in increased temperature on the ground.[7] On the other hand, increased temperatures can push wildfires, increase consumption,[8] and cause "stagnation events" that result in hotter air pollutants to become trapped on the ground rather than circulate in the atmosphere.[9] All of this has the potential to substantially degrade air quality and pose a serious threat to human health, and as such, it is worth further investigation.

The relationship between air quality and temperature will be explored through the **Air Quality Data Set**, which was donated to the University of California Irvine (UCI) Machine Learning Repository by the *Italian National Agency for New Technologies, Energy and Sustainable Economic Development*.[10] The dataset contains responses from metal oxide sensors that were located in a polluted area, at road level, within an Italian city. The data was

---

[1] "How Is Air Quality Measured? | Noaa Scijinks – All about Weather," https://scijinks.gov/air-quality/, (accessed May 28, 2022).
[2] "Sustainable Development Goals & Air Pollution," https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/policy-progress/sustainable-development-goals-air-pollution, (accessed May 28, 2022).
[3] "Air Pollution," https://www.who.int/health-topics/air-pollution, (accessed May 28, 2022).
[4] "Sustainable Development Goals & Air Pollution"
[5] "Air Pollution"
[6] Centers for Disease Control and Prevention, *Climate Change Decreases the Quality of the Air We Breathe*, https://www.cdc.gov/climateandhealth/pubs/air-quality-final_508.pdf, (accessed May 28, 2022).
[7] European Union, *Air Pollution and Climate Change*, https://ec.europa.eu/environment/integration/research/newsalert/pdf/24si_en.pdf, (accessed May 28, 2022).
[8] CDC, *Climate Change Decreases the Quality of the Air We Breathe*.
[9] "Climate Change Is Threatening Air Quality across the Country," https://www.climatecentral.org/news/ climate-change-is- threatening-air-quality-across-the-country-2019, (accessed May 28, 2022).
[10] UCI Machine Learning Repository, *Air Quality Data Set*, https://archive.ics.uci.edu/ml/datasets/Air+quality, (accessed May 26, 2022).

recorded hourly from March 10, 2004 to April 4, 2005, and it was compared against independent measurements by a co-located reference certified analyzer (hereafter referred to as "ground truth"). Each row in the dataset contains values for the following:

- date,
- time,
- measured and ground truth concentrations of carbon monoxide (CO) in mg/m^3,
- measured and ground truth concentrations of non-methane hydrocarbons (NMHC) in $\mu g/m^3$,
- ground truth concentrations of benzene (C_6H_6) in $\mu g/m^3$,
- measured and ground truth concentrations of nitric oxide (NO_x) in parts per billion (ppb),
- measured and ground truth concentration of nitrogen dioxide (NO_2) in $\mu g/m^3$,
- measured concentration for ozone gas (O_3),
- absolute humidity,
- relative humidity, and
- temperature in degrees Celsius.

Some preparation on the dataset was required before proper statistical analysis could be conducted. First, the hourly data was rolled-up to daily data by taking the daily average (ignoring missing data). Second, columns with substantial amounts of missing data (after taking the daily average) were removed: ground truth concentrations of CO, NMHC, NO_x, and NO_2. Furthermore, as humidity is already known to be inversely proportional with temperature,[11] the columns related to humidity were also removed to more carefully explore the effect of air quality on temperature. Finally the remaining missing values (about 8 days) were imputed using the global mean of their respective columns. This leaves measurements for CO, C_6H_6, NMHC, NO_x, NO_2 and O_3 to predict temperature.

[11] De-Zheng Sun and Abraham H. Oort, "Humidity–Temperature Relationships in the Tropical Troposphere," *Journal of Climate* 8, no. 8 (1995): 1974–87, http://www.jstor.org/stable/26200032.; "Chapter 7 - Relationship between Temperature and Moisture | Animal & Food Sciences," https://afs.ca.uky.edu/poultry/chapter-7-relationship-between-temperature-and-moisture, (accessed May 28, 2022).

Analysis and Interpretation

An initial survey of the dataset in Figure 1 shows no obvious patterns among the predictors. However, a slight increase in temperature can be seen from June to September, which is to be expected as this corresponds with the summer months in Italy. This may indicate yearly seasonality in the data; however, since the data only spans one year, the seasonality is not given a chance to present itself.

FIGURE 1. *Time series plot of the six predictors and the response variable (temperature).*

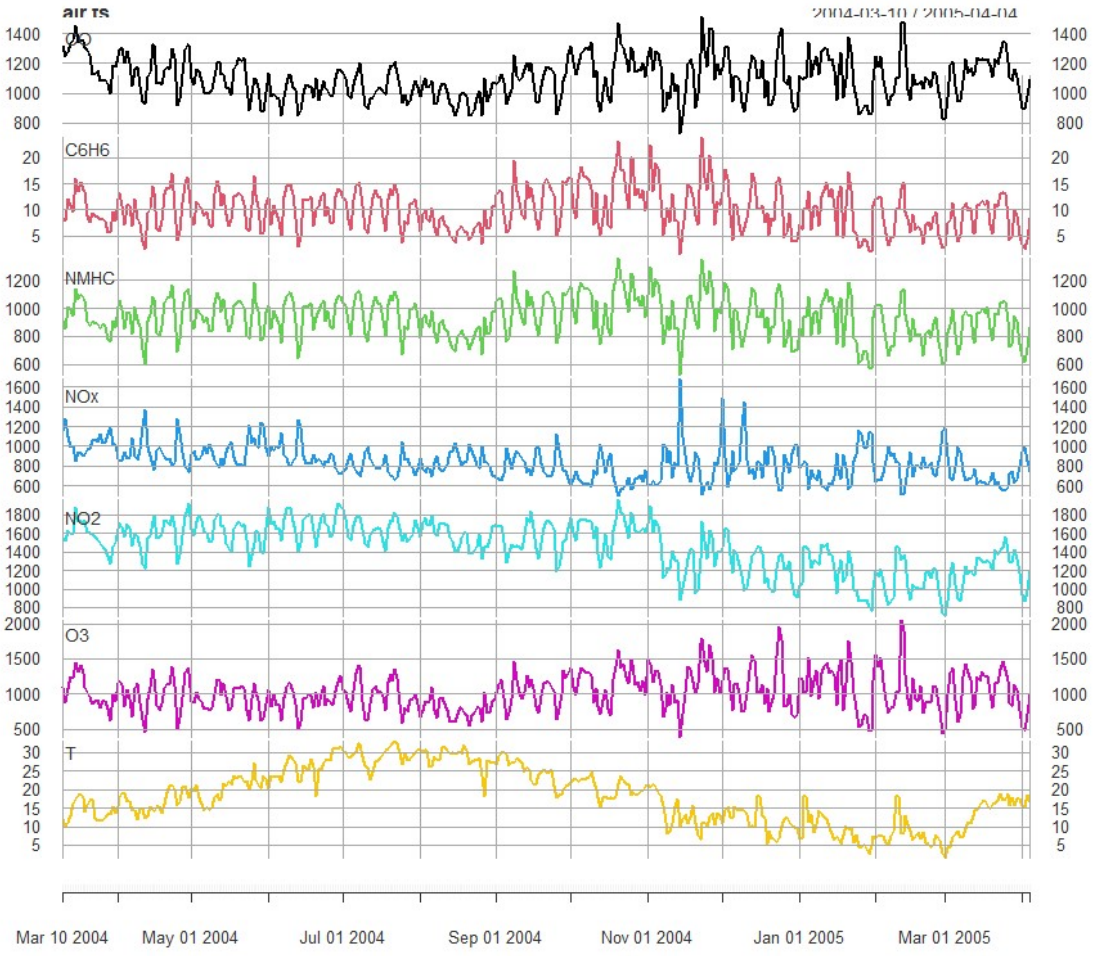
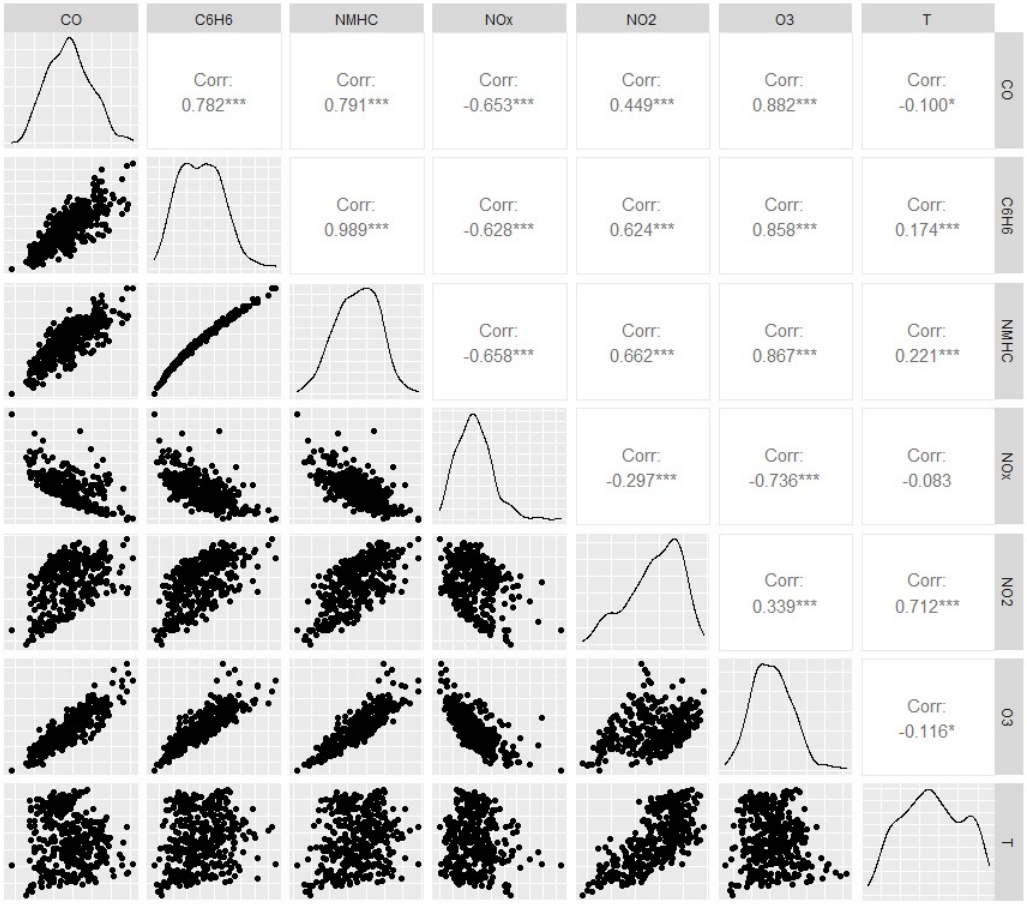


Figure 2 is a scatterplot matrix of seven (7) variables — the six predictors mentioned above and response variable (temperature). The last row shows the relationships between the response variable and each of the predictors. The strength of these relationships are shown by the correlation coefficients along the first column. The remaining scatterplots and correlation coefficients show the relationship between the predictors.

FIGURE 2. *Scatterplot matrix the temperature (T) and the six predictors.*



Notably, Figure 2 shows that C_6H_6 is highly correlated with NMHC with a correlation coefficient of 0.989. This suggests that models that include both C_6H_6 and NMHC as predictors may have high levels of multicollinearity. As benzene is also a non-methane hydrocarbon, it makes sense that high amounts of benzene would be found where high amounts of NMHCs are found. Moreover, Figure 2 also shows that O_3 is highly correlated with CO, C_6H_6, NMHC, and NO_x with correlation coefficients of 0.882, 0.858, 0.867, and -0.736, respectively. This also suggests that models that include these predictors together may have higher levels of multicollinearity. The high negative correlation NO_x and O_3 may be because ground-level ozone is a byproduct of many of the chemical reactions that take place between pollutants, such as NO_x and volatile organic compounds.[12] This means that NO_x tends to turn into O_3.

[12] Minnesota Pollution Control Agency, "Ozone," January 24, 2013, https://www.pca.state.mn.us/air/ozone, (accessed May 28, 2022).

The *regsubsets* function from the *leaps* R package was used on the dataset to generate six (6) potential linear regression models, one for each possible dimension of the model, determined by the number of predictors selected (i.e. 1 predictor, 2 predictors, and so on, up to 6 predictors). The *regsubsets* function works by automatically selecting the model with the combination of predictors that gives the highest Adjusted R^2 for each dimension. Adjusted R^2 indicates how well data points fit a curve or line while penalizing additional insignificant predictors. It indicates the proportion of the variance in the response variable (temperature) that was captured by the predictors.

To compare the models against each other, the Variance Inflation Factor (VIF) of each model was also computed. VIF detects multicollinearity in regression analysis, and multicollinearity occurs when there is a correlation between variables in a model, and its presence can have adverse effects on the results of the regression. High levels of multicollinearity reduce interpretability between variables and affect the reliability of predictions. Thus, VIF estimates how much the variance is inflated because of multicollinearity. A VIF of 1 indicates no correlation, a VIF between 1 and 5 indicates a moderate correlation, and a VIF greater than 5 indicates a high correlation. The best model would then be selected as the model with the highest Adjusted R^2 and where each of its predictors had an acceptable VIF (≤ 5). Note that VIF was not computed for the one-dimensional model since this only had a single predictor and thus could not be affected by multicollinearity. Other relevant indicators such as BIC and Mallow's CP were also included. The six (6) generated models for the dataset are summarized in Table 1 below.

TABLE 1. *Summary of generated models from the cleaned and imputed Air Quality Dataset using the regsubsets function.*

Model Name	Dimension	Predictors Included	Adjusted R^2	BIC	Mallows' CP	Highest VIF
mod1	1	NO_2	0.5059691	-264.7829	435.062136	NA
mod2	2	CO, NO_2	0.7258193	-490.0457	70.060879	1.251875

mod3	3	CO, NO_x, NO_2	0.7619788	-540.3841	11.028512	1.991629
mod4	4	CO, C_6H_6, NO_x, NO_2	0.7661351	-542.3150	5.147589	3.677295
mod5	5	CO, C_6H_6, NMHC, NO_x, NO_2	0.7668302	-538.5244	5.002469	68.337538
mod6	6	CO, C_6H_6, NMHC, NO_x, NO_2, O_3	0.7662245	-532.5582	7.000000	86.637864

Looking at the VIF for each model indicates that **mod5** and **mod6** are highly multicollinear, rendering them unsuitable for linear regression. This aligns with the group's preliminary analysis of the data as **mod6** includes O_3 and **mod5** includes both C_6H_6 and NMHC. As shown in Figure 2 and as discussed above, these predictors are highly correlated. Among the remaining models, **mod4** has the best results with an Adjusted R^2 of 0.7661. This indicates that 76.61% of the variation in temperature was explained by the predictors. This model also had a BIC of -542.3150 and Mallows' CP of 5.1476. In fact, **mod4** has the best (lowest) BIC among all the models. The predictors of **mod4** and the respective estimates for their coefficients are summarized in the Table 2 below.

TABLE 2. *Predictors of the best linear regression model with 4 predictors.*

Predictor	**Coefficient**	**P-Value**
(Intercept)	28.124923	2×10^{-16}
CO	-0.034868	2×10^{-16}
C_6H_6	-0.253247	0.00526
NO_x	-0.012821	1.97×10^{-15}

NO_2	0.028734	2×10^{-16}

Some insights can be gleaned from the values of these coefficients. First, just looking at the coefficients may indicate that CO has the second strongest influence on temperature, but in actuality, it has the least effect per unit of concentration as CO is measured in milligrams while the rest are measured in micrograms. This indicates that it actually has the least influence as it requires changes a thousand times more potent (a milligram is a thousand micrograms) to get an effect on temperature in the same order as NO_x and NO_2.

Second, the absolute value of the coefficient of C_6H_6 (0.253247) is an order of magnitude larger than the rest of the coefficients. This indicates that benzene has the largest influence on temperature according to the model. That is, a one unit change in benzene would result in the largest change in temperature.

Third, the coefficients for the predictors of CO, C_6H_6, and NO_x are negative, indicating a negative relationship between these predictors and temperature. Interestingly, this seems to contradict what was mentioned in the background, i.e., that such emissions contribute to global warming and thus a positive relationship between these and temperature is expected. This does not suggest, however, that the well-known and empirically established relationship between emissions and climate change is false. Instead, there are several reasons why this inverse relationship arose from the regression.

Firstly, as established in the discussion of Figure 1, temperature varies depending on the time of year, coinciding with the summer and winter months of Italy. This means that the natural shifts in temperature due to changing seasons could outweigh the effects of emissions within the same year. It would be better to compare the same seasons across multiple years in order to identify global warming, however this was not possible in this case since the dataset only spanned one year.

Secondly, the chemicals may have coolant properties or are heavily correlated to other chemicals with coolant properties. Both C_6H_6 and NO_x have industrial uses as coolants due to their high thermal capacity and low viscosity,[13] while CO easily reacts with free oxygen in the air to become CO_2, which is a coolant.[14]

[13] Compilation of Organic Moderator and Coolant Technology. United States Atomic Energy Commission, Technical Information Service Extension, 1957

[14] "Carbon Dioxide as a Coolant: Developments in the Application of Liquid Carbon Dioxide to Machining Operations." Aircraft Engineering and Aerospace Technology 26, no. 7 (January 1, 1954): 234–234. https://doi.org/10.1108/eb032448; "Effects of CO2 in Air on PH of Ethylene Glycol Based Coolant." Accessed

Third and lastly, it is likely that there are other ways in which pollution levels and temperature are related which would explain the inverse relationship. For one, vehicle emissions (which likely make up most of the emissions in the dataset since collection was done on a thoroughfare) could increase in the cold because low temperatures decrease the efficiency of automobile engines, causing them to burn more fuel.[15] For another, ground-level air pollution tends to dissipate faster in warmer weather due to rising warm air near the ground lifting pollution away.[16]

However, on the other hand, the coefficient NO_2 has a positive relationship with temperature. While this is the expected relationship based on the literature covered in the background, this relationship is now the outlier among the other parameters in this model. There are some possible reasons for why this is the case. NO_2 is formed naturally in warm conditions in water and may become common in warmer periods of the year.[17] NO_2 forms anthropogenically from the burning of fuel specifically at high heat which may distinguish it from the other variables by exclusively being emitted in periods of high temperature.[18] This may also be caused by some reason unknown to the authors and will need further investigation.

When using a linear regression model, a few assumptions are made implicitly. As such, in order to guarantee the reliability of its predictions, these assumptions must hold true.

First, the residuals must be unrelated to the predictor variables; otherwise, this suggests that not all of the information was captured by the model. The residuals from the best linear regression model with 4 predictors (**mod4)** plotted against each predictor in Figure 3 indicate no obvious patterns and seem to be randomly scattered. This suggests that the assumption is satisfied.

Second, the residuals must be unrelated to the predictions (fitted values) made by the model; otherwise, this may exhibit heteroscedasticity, which indicates that the variance of the residuals may not be constant. The residuals from **mod4** plotted against fitted values can be found below in Figure 4. The figure shows no obvious relationship between residuals and the

---

May 28, 2022. https://www.vanchem.com/rockwell-thermal-fluids/coolants/effects-of-co2-in-air-on-ph-of-ethylene-glycol-based-coolant/.

[15] United States Environmental Protection Agency, "Fuel Economy in Cold Weather," https://www.fueleconomy.gov/feg/coldweather.shtml, (accessed May 28, 2022).

[16] University Corporation for Atmospheric Research, "How Weather Affects Air Quality," https://scied.ucar.edu/learning-zone/air-quality/how-weather-affects-air-quality, (accessed May 28, 2022).

[17] Chernikov, Anatoly V., Vadim I. Bruskov, and Sergey V. Gudkov. "Heat-Induced Formation of Nitrogen Oxides in Water." Journal of Biological Physics 39, no. 4 (September 2013): 687–99. https://doi.org/10.1007/s10867-013-9330-z.

[18] Minnesota Pollution Control Agency. "Nitrogen Dioxide (NO2)," April 4, 2017. https://www.pca.state.mn.us/air/nitrogen-dioxide-no2.

fitted values. Furthermore, running a non-constant variance score test provides a p-value of 0.31141, suggesting that the model follows the assumption of constant variance.

FIGURE 3. *Scatterplots of the residuals versus each predictor.*

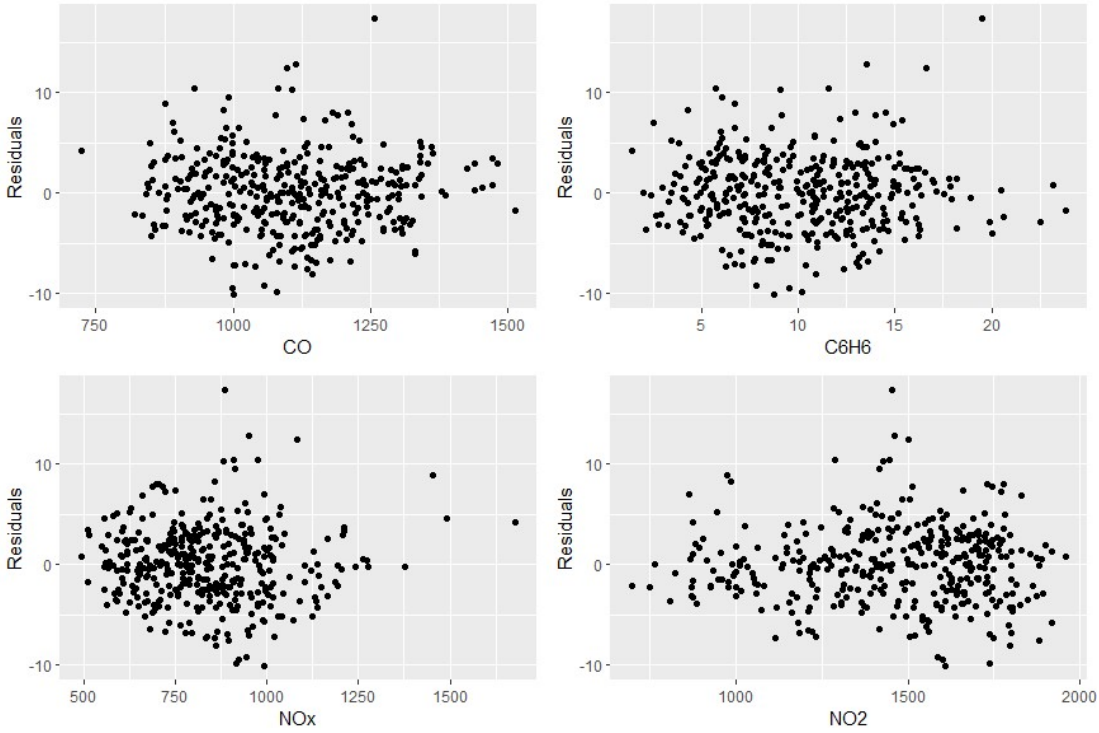
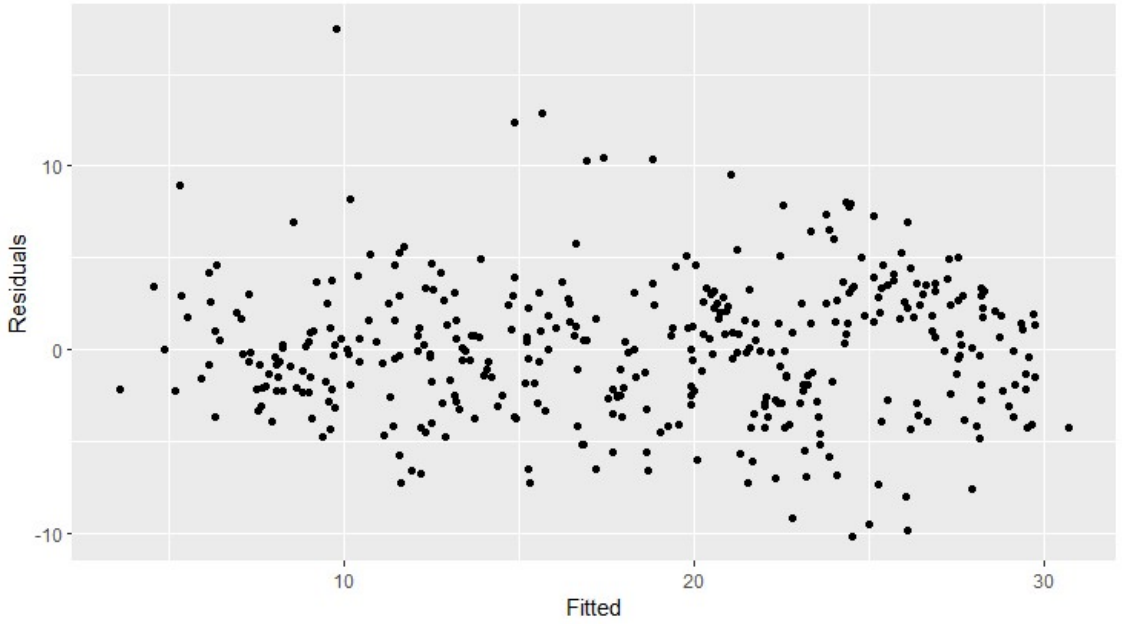
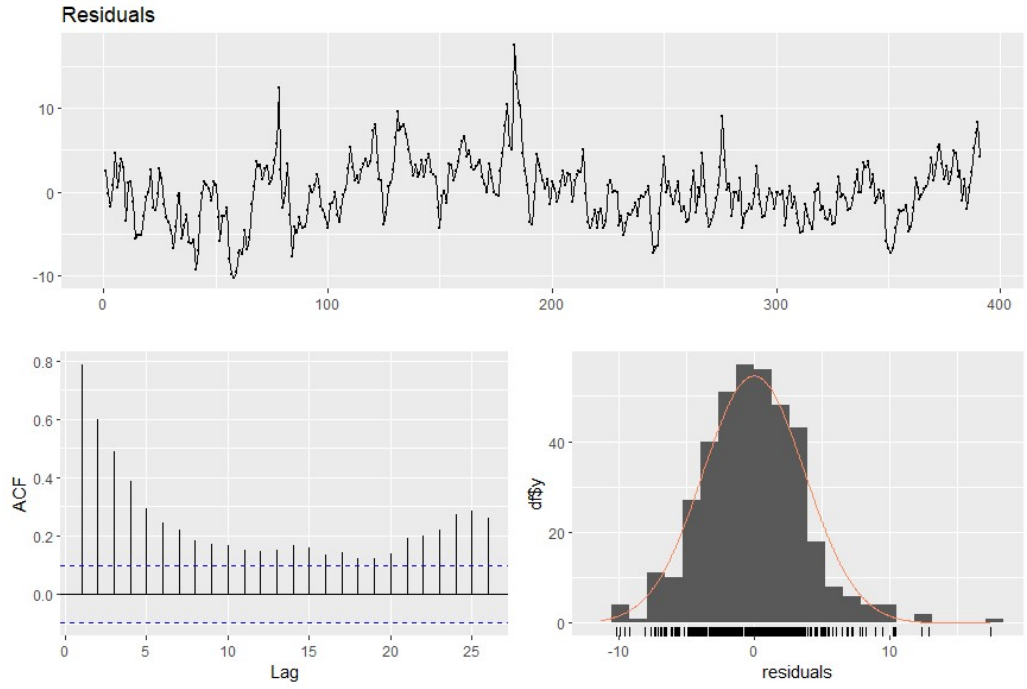


FIGURE 4. *Scatterplot of the residuals versus fitted values.*



Third, the residuals must have a mean of zero; otherwise the predictions show some systematic bias to an outcome for one reason or another. The line graph and histogram of the residuals in Figure 5 suggest that the residuals have a mean of zero, confirming the assumption.

FIGURE 5. *Residuals plots of the best linear regression model with 4 predictors.*

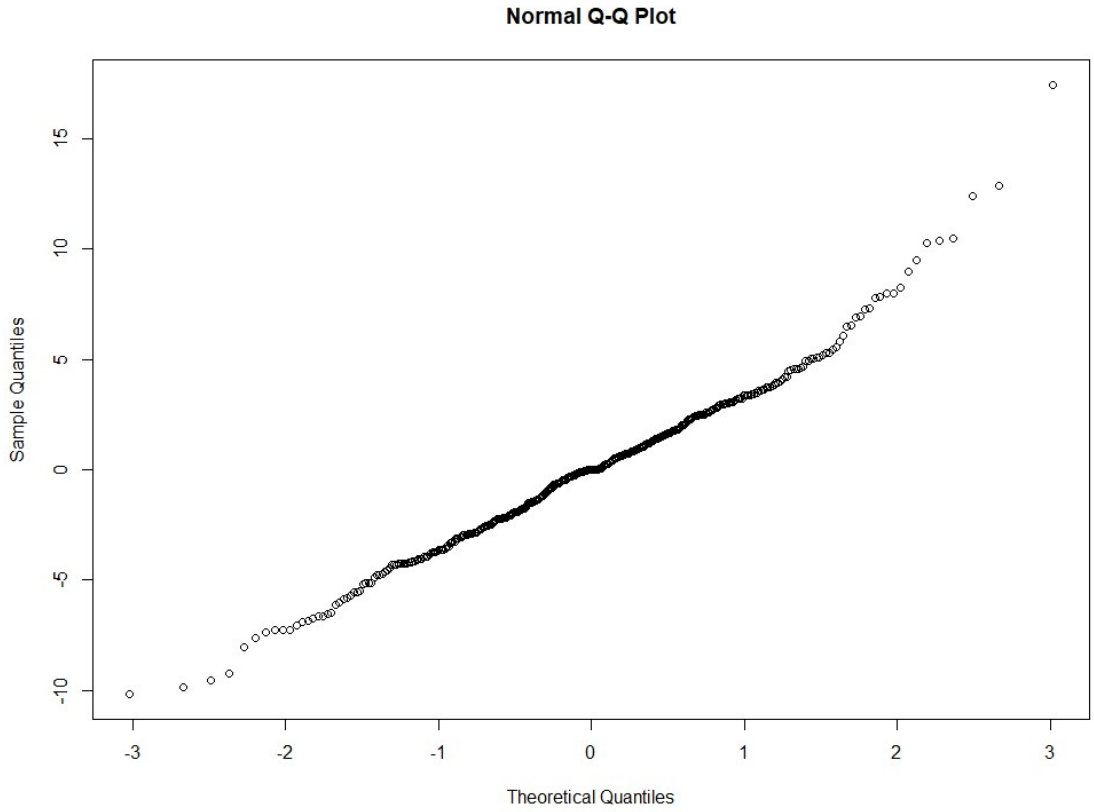


Notably, looking at the time series plot of the residuals in Figure 5, it no longer has the yearly seasonality (i.e., the slight increase from June to September) that was characteristic of the response variable (temperature). This suggests that the linear regression model was able to capture the seasonality in the temperature. This may be because the use of vehicles (the primary source of the predictors) are also seasonal in nature, i.e., when people tend to use more vehicles during winter and when vehicles burn fuel less effectively during winter.

Additionally, the residuals have to be normally distributed to make the calculation of prediction intervals easier. The histogram in Figure 5 above seems to indicate that the residuals are approximately normally distributed, but a closer look at the Q-Q plot in Figure 6 below and an application of the Shapiro-Wilk normality test (with p-value of 0.0001419) show that the residuals are not actually normally distributed. Hence, this suggests that one of some possible outcomes: some extreme observations are seriously skewing parameter estimates, a better model may be available, there are some unusual data points that should be examined more closely, or there are some other problems with the model assumptions.[19]

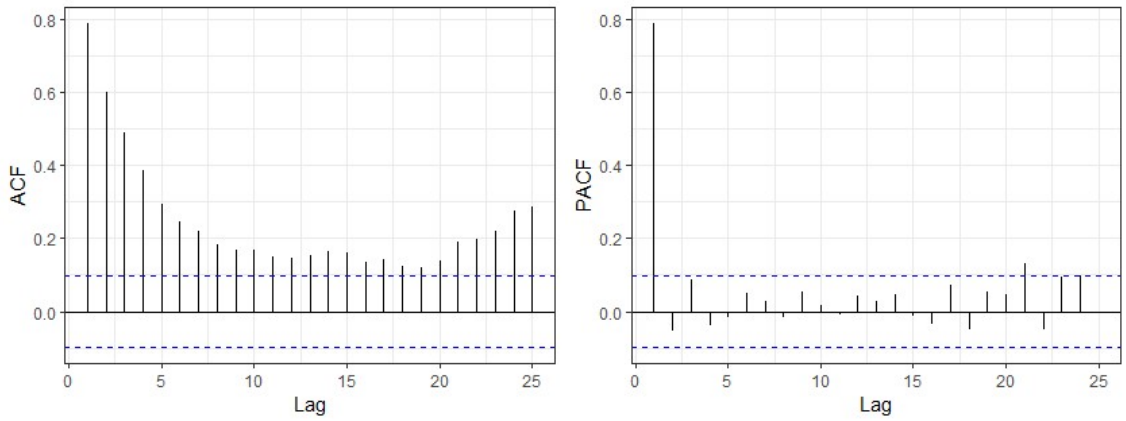
[19] "Testing the Assumptions of Linear Regression." Accessed May 28, 2022. https://people.duke.edu/~rnau/testing.htm#

FIGURE 6. *Normal Q-Q Plot of the residuals of the best linear regression model with 4 predictors.*



Another assumption in linear regression is that the residuals are not autocorrelated; otherwise, the predictions will be inefficient as more information in the data can be exploited. The plots of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) with Bartlett's band in Figure 7 shows extensive autocorrelation. The lag shows rapid decay, possibly indicating stationarity, but remaining above Bartlett's band. As the regression was done over time series data, some autocorrelation was to be expected. Note that predictions from models with autocorrelated residuals are still unbiased (from the assumption the residuals have mean zero), but these models would usually have larger prediction intervals compared to models that do not have autocorrelated residuals.

FIGURE 7. *Plots of ACF and PACF of the residuals of the best linear regression model with 4 predictors.*



Since the residuals of **mod4** (hereafter referred to as "errors") were determined to be autocorrelated, the next task is to model this autocorrelation with an autoregressive-moving-average (ARMA) model. The Extended ACF (EACF) table in Figure 8 was used to determine the order of the ARMA model.

FIGURE 8. *Sample Extended ACF Table for the errors.*

AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	x	x	x	x	x	x	x	x	x	x	x	x	x
1	o	x	o	o	o	o	o	o	o	o	o	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	o
3	x	x	o	o	o	o	o	o	o	o	o	o	o
4	x	x	o	o	o	o	o	o	o	o	o	o	o
5	x	x	x	x	o	o	o	o	o	o	o	o	o
6	x	x	x	x	x	o	o	o	o	o	o	o	o
7	x	x	x	o	x	x	o	o	o	o	o	o	o
8	x	x	x	o	x	x	x	o	o	o	o	o	o
9	x	x	x	x	x	o	o	o	o	o	o	o	o
10	x	o	x	x	x	x	o	o	x	o	o	o	o
11	x	x	x	x	x	x	o	o	x	o	o	o	o
12	x	x	x	o	x	o	o	x	o	o	o	o	o

The EACF table indicates some promising candidates to model, namely **ARMA(2,1)**, **AR(1)**, **ARMA(1,1)**, **ARMA(1,2)**, and **ARMA(2,2)**, all entries at the tip of the "triangle".

After converting the residuals into a time series object, the residuals were fitted using the ARMA function with orders corresponding to the candidates identified above. The models and their respective AICs can be found in Table 3 below.

TABLE 3. *AIC of the Autoregressive-Moving-Average (ARMA) models for the errors.*

Model	AIC
AR(1)	1770.79
ARMA(1,1)	1771.52
ARMA(1,2)	1770.97
ARMA(2,1)	1763.25
ARMA(2,2)	1772.95

From Table 3, the model with the lowest AIC was **ARMA(2,1)** with an AIC of 1763.25. Thus, this was chosen as the candidate model for the errors. Table 4 shows the coefficients of the **ARMA(2,1)** model.

TABLE 4. *Coefficients of the parameters of the **ARMA(2,1)** model.*

Parameter	Coefficient	P-Value
Intercept	-0.0025	0.99129
ar_1	-0.1237	0.00109
ar_2	0.7077	2×10^{-16}
ma_1	0.9719	2×10^{-16}
Intercept	-0.0025	0.99129

Relevantly, the p-value for the intercept is very high, indicating that it is not significantly different from zero. This satisfies the implicit assumption in linear regression that the mean of the errors are zero.

Combining these coefficients with the best regression model (**mod4**), the following is the final fitted model:

$$y_t = 28.1249 - 0.0349x_{CO} - 0.2532x_{C_6H_6} - 0.0128x_{NO_x} + 0.02873x_{NO_2} + e_t$$

$$e_t = -0.0025 - 0.1237e_{t-1} + 0.7077e_{t-2} + a_t + 0.9719a_{t-1}$$

$$a_t \sim WN(0, 5.241)$$

To check the validity of the ARMA model, further analysis was performed on its residuals.

FIGURE 9. *ACF and PACF plots of the residuals of the* ***ARMA(2,1)*** *model.*

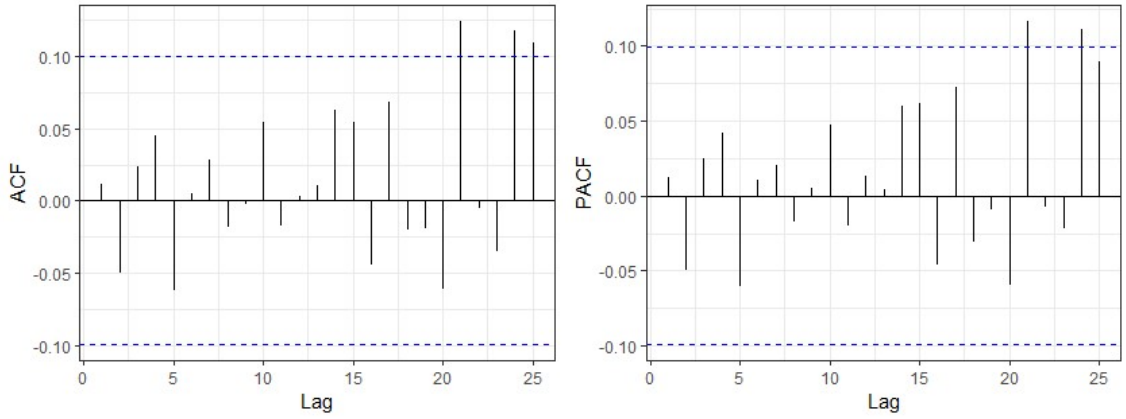


Figure 9 above shows the ACF and PACF plots of the residuals of the error term under the **ARMA(2,1)** model. Here, it can be seen that the values fall within the Barlett's band for small lags, suggesting that these residuals are not autocorrelated. More testing for autocorrelation was then performed using the Ljung-Box test up to lag 6 as suggested by the log-length rule of thumb as it generally provides the highest power. As the p-value is 0.7382, the null hypothesis (i.e., that the ACF values up to lag 6 are all equal to 0) failed to be rejected. This again suggests that there is **no remaining autocorrelation in the residuals** and the **ARMA(2,1)** model has successfully captured the remaining time-variant trends.

Next, comparing both the fitted values from the regression model and the fitted values from the regression model with ARMA errors against temperature over time shows that the regression model with ARMA errors is more accurate. These graphs can be found below in Figures 10 and 11.

Retrieving their root mean squared error (RMSE) confirms that the addition of **ARMA(2,1)** errors to the model indeed improves performance, going from an RMSE of 3.7646 to 2.2804. A change of around 1.4 RMSE is quite significant, as the predicted values are in

units of Celsius, where even a 1 degree change can have large environmental effects over the span of a year.[20]

FIGURE 10. *Line graph of the fitted values from the best regression model* ***(mod4)*** *against temperature over time.*

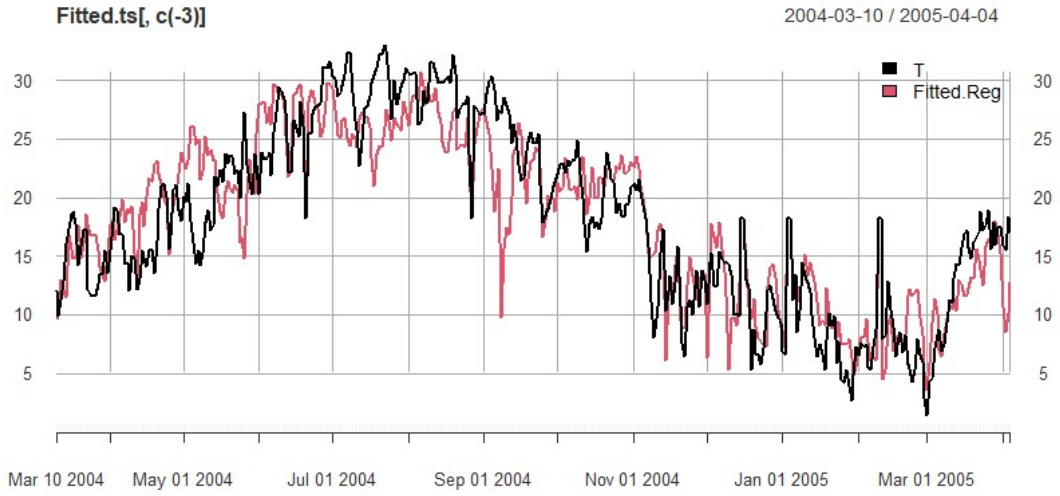


FIGURE 11. *Line graph of the fitted values from the best regression model* ***(mod4)*** *with* ***ARMA(2,1)*** *errors against temperature over time.*

[20] National Aeronautics and Space Association, "A Degree of Concern: Why Global Temperatures Matter." Climate Change: Vital Signs of the Planet. Accessed May 28, 2022. https://climate.nasa.gov/news/2865/a-degree-of-concern-why-global-temperatures-matter; Environmental Defense Fund. "How Can Half a Degree of Warming Matter so Much?" Accessed May 28, 2022. https://www.edf.org/blog/2018/10/18/how-can-half-degree-warming-matter-so-much.

As mentioned, the model is successful in capturing some time-variant trends in the dataset. There are a few reasons for why this might be the case. The first and most obvious is that temperature and air conditions can persist in an area for more than a few days, lingering or dissipating over time. Second, temperature is also heavily influenced by broader weather conditions and the seasons, which as mentioned earlier, may be reflected in the rise in temperature during the summer months of Italy. Due to the year-long timespan of the dataset, while longer-term trends such as global warming cannot be captured, the changing of seasons will reflect in the change in time. Third and lastly, shorter term trends such as weekends and school schedules may affect traffic and pollution outcomes, indirectly affecting temperature as a second-round effect.

Notably, there seems to be large discrepancies in predicted and actual values at the beginning of September 2004. Taking a closer look at the models as well as the dataset, the authors determined that this discrepancy occurred around September 8, 2004 due an large increase in detected benzene (C_6H_6). Interestingly, this date coincides with the eruption of Mt. Etna, an active stratovolcano located in Italy.[21] Volcanoes are known to produce benzene,[22] which may have been detected by the instruments used to collect the *Air Quality Dataset*. As such, the large increase in benzene in the atmosphere caused the fitted model to wrongly predict a large drop in temperature.

Conclusion

While the initial direction of the background of the paper and the intuition of the authors might lead to the conclusion that increases in air pollution lead to increases in temperature, the dataset and the results show something quite different. Given the results of multiple linear regression analysis, the authors find that over the span of a year in Italy, increases in CO, C_6H_6, and NO_x correlate negatively with regional temperature, while NO_2 correlates positively with regional temperature. The predictive capabilities of the model improve substantially when the error term is modeled with **ARMA(2, 1)**, as the dataset exhibits autocorrelation. Contrary to intuition, air pollutants may increase when temperatures are low due to the increased consumption of fuel or the short-term coolant effects of some chemicals. Future studies may want to examine the data over a longer time-period to capture changes in temperature from

[21] "Global Volcanism Program | Report on Etna (Italy) — September 2004," https://volcano.si.edu/ShowReport.cfm?doi=10.5479/si.GVP.BGVN200409-211060, (accessed May 28, 2022).
[22] Delaware Health and Social Services, *Benzene*, https://dhss.delaware.gov/dph/files/benzenefaq.pdf, (accessed May 28, 2022).

long-term trends like global warming or urbanization. Future studies may also want to examine the usefulness of a model incorporating seasonality such as **SARMA** to capture seasonal fluctuations in temperature due to the changing of seasons, or more granular weekly (or even hourly) trends.

References

"Air Pollution." Accessed May 28, 2022. https://www.who.int/health-topics/air-pollution.

Centers for Disease Control and Prevention. *Climate Change Decreases the Quality of the Air We Breathe*. Accessed May 28, 2022. https://www.cdc.gov/climateandhealth/pubs/air-quality-final _508.pdf.

"Chapter 7 - Relationship between Temperature and Moisture | Animal & Food Sciences." Accessed May 28, 2022. https://afs.ca.uky.edu/poultry/chapter-7-relationship- between-temperature-and-moisture.

"Climate Change Is Threatening Air Quality across the Country." Accessed May 28, 2022. https://www.climatecentral.org/news/climate-change-is-threatening-air-quality-across-the-country-2019.

"Carbon Dioxide as a Coolant: Developments in the Application of Liquid Carbon Dioxide to Machining Operations." Aircraft Engineering and Aerospace Technology 26, no. 7 (January 1, 1954): 234–234. https://doi.org/10.1108/eb032448.

"Effects of CO2 in Air on PH of Ethylene Glycol Based Coolant." Accessed May 28, 2022. https://www.vanchem.com/rockwell-thermal-fluids/coolants/effects-of-co2-in-air-on-ph-of-ethylene-glycol-based-coolant/.

"Testing the Assumptions of Linear Regression." Accessed May 28, 2022. https://people.duke.edu/~rnau/testing.htm.

Chernikov, Anatoly V., Vadim I. Bruskov, and Sergey V. Gudkov. "Heat-Induced Formation of Nitrogen Oxides in Water." Journal of Biological Physics 39, no. 4 (September 2013): 687–99. https://doi.org/10.1007/s10867-013-9330-z.

Compilation of Organic Moderator and Coolant Technology. United States Atomic Energy Commission, Technical Information Service Extension, 1957.

Delaware Health and Social Services. *Benzene*. Accessed May 28, 2022. https://dhss.delaware.gov/dph/files/ benzenefaq.pdf.

Environmental Defense Fund. "How Can Half a Degree of Warming Matter so Much?" Accessed May 28, 2022. https://www.edf.org/blog/2018/10/18/how-can-half-degree-warming-matter-so-much.

European Union. *Air Pollution and Climate Change*. Accessed May 28, 2022. https://ec.europa.eu/environment/integration/research/newsalert/pdf/24si_en.pdf.

"Global Volcanism Program | Report on Etna (Italy)." Accessed May 28, 2022. https://volcano.si.edu/ShowReport.cfm?doi=10.5479/si.GVP.BGVN200409-211060.

"How Is Air Quality Measured? | Noaa Scijinks – All about Weather." Accessed May 28, 2022. https://scijinks.gov/air-quality/.

Minnesota Pollution Control Agency. "Ozone." January 24, 2013. https://www.pca.state.mn.us/air/ozone.

Minnesota Pollution Control Agency. "Nitrogen Dioxide (NO2)," April 4, 2017. https://www.pca.state.mn.us/air/nitrogen-dioxide-no2.

National Aeronautics and Space Association, "A Degree of Concern: Why Global Temperatures Matter." Climate Change: Vital Signs of the Planet. Accessed May 28, 2022. https://climate.nasa.gov/news/2865/a-degree-of-concern-why-global-temperatures-matter;

Sun, De-Zheng, and Abraham H. Oort. "Humidity–Temperature Relationships in the Tropical Troposphere." *Journal of Climate* 8, no. 8 (1995): 1974–87. http://www.jstor.org/stable/26200032.

"Sustainable Development Goals & Air Pollution" Accessed May 28, 2022. https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/policy-progress/sustainable-development-goals-air-pollution.

UCI Machine Learning Repository. *Air Quality Data Set*. Accessed May 26, 2022. https://archive.ics.uci.edu/ml/datasets/Air+quality.

United States Environmental Protection Agency. 'Fuel Economy in Cold Weather'. Accessed 28 May 2022. https://www.fueleconomy.gov/feg/coldweather.shtml.

University Corporation for Atmospheric Research. 'How Weather Affects Air Quality'. Accessed 28 May 2022. https://scied.ucar.edu/learning-zone/air-quality/how-weather-affects-air-quality.

Appendix

R Code

```
library(tseries)
library(TSA)
library(car)
library(leaps)
library(tidyverse)
library(lubridate)
library(zoo)
library(xts)
library(quantmod)
library(forecast)
library(gridExtra)
library(ggplot2)
library(fBasics)

air = read.csv('imputed_airquality.csv')

#####################################################################
# VISUALIZATIONS

air %>% select(c(-1,-9)) %>% GGally::ggpairs() + theme(axis.line=element_blank(),
                                                       axis.text=element_blank(),
                                                       axis.ticks=element_blank())
air.ts = xts(air[c(-1,-9)], order.by = as.Date(air$Date))
plot.xts(air.ts, multi.panel=TRUE, yaxis.same=FALSE)

acf(air.ts[,c('T')], lag.max=365)

#####################################################################
# REGRESSION

models = regsubsets(T~., data=air.ts, method="exhaustive")
```

```
summary(models)

AdjR2 = summary(models)$adjr2
Mallows.cp = summary(models)$cp
BIC = summary(models)$bic
Models = 1:6
Results = cbind(Models, AdjR2, Mallows.cp, BIC)
Results # Best BIC is model 4, Best AdjR2 is model 5

mod6 = lm(T~CO+C6H6+NMHC+NOx+NO2+O3, data=air.ts)
mod5 = lm(T~CO+C6H6+NMHC+NOx+NO2, data=air.ts)
mod4 = lm(T~CO+C6H6+NOx+NO2, data=air.ts)
mod3 = lm(T~CO+NOx+NO2, data=air.ts)
mod2 = lm(T~CO+NO2, data=air.ts)
mod1 = lm(T~NO2, data=air.ts)

vif(mod6)
vif(mod5)
vif(mod4) # Looks good. Highest is 3.677
vif(mod3)
vif(mod2)

summary(mod4)

#########################################################################
# REGRESSION RESIDUAL ANALYSIS

e = residuals(mod4)
checkresiduals(mod4)

regdf = as.data.frame(air.ts)
regdf['Residuals'] = as.numeric(e)
```

```
p1 = ggplot(regdf, aes(x=CO, y=Residuals)) + geom_point()
p2 = ggplot(regdf, aes(x=C6H6, y=Residuals)) + geom_point()
p3 = ggplot(regdf, aes(x=NOx, y=Residuals)) + geom_point()
p4 = ggplot(regdf, aes(x=NO2, y=Residuals)) + geom_point()
gridExtra::grid.arrange(p1,p2,p3,p4, nrow=2)

cbind(Fitted = mod4$fitted.values, Residuals=e) %>% as.data.frame() %>%
  ggplot(aes(x=Fitted,y=Residuals)) + geom_point()
as.numeric(e)
ncvTest(mod4)
qqnorm(e)
shapiro.test(e) # We fail the shapiro test

#########################################################################
# ARMA ON RESIDUAL

e.ts = xts(e, order.by = as.Date(air$Date))
plot.xts(e.ts)
acf(e.ts)
pacf(e.ts, lag.max=100)

ggacf <- function(x, ci=0.95, type="correlation", xlab="Lag", ylab=NULL, ylim=NULL,
main=NULL, ci.col="blue", lag.max=NULL) {
  x <- as.data.frame(x)
  x.acf <- acf(x, plot=F, lag.max=lag.max, type=type)
  ci.line <- qnorm((1 - ci) / 2) / sqrt(x.acf$n.used)
  d.acf <- data.frame(lag=x.acf$lag, acf=x.acf$acf)
  g <- ggplot(d.acf, aes(x=lag, y=acf)) +
    geom_hline(yintercept=0) +
    geom_segment(aes(xend=lag, yend=0)) +
    geom_hline(yintercept=ci.line, color=ci.col, linetype="dashed") +
    geom_hline(yintercept=-ci.line, color=ci.col, linetype="dashed") +
    theme_bw() +
```

```
  xlab("Lag") +
  ggtitle(ifelse(is.null(main), "", main)) +
  if (is.null(ylab))
    ylab(ifelse(type=="partial", "PACF", "ACF"))
 else
  ylab(ylab)

 g
}

grid.arrange(ggacf(e.ts), ggacf(e.ts, type="partial"), ncol=2)

eacf(e.ts, ar.max=12, ma.max=12)
# Suggest ARMA(2,1), but let's try AR(1), ARMA(1,1), ARMA(1,2), ARMA(2,2)

include = TRUE
tsmod1 = arma(e.ts, order=c(2,1), include.intercept = include)
tsmod2 = arma(e.ts, order=c(1,0), include.intercept = include)
tsmod3 = arma(e.ts, order=c(1,1), include.intercept = include)
tsmod4 = arma(e.ts, order=c(1,2), include.intercept = include)
tsmod5 = arma(e.ts, order=c(2,2), include.intercept = include)

summary(tsmod1) # Looks like best is ARMA(2,1)
summary(tsmod2)
summary(tsmod3)
summary(tsmod4)
summary(tsmod5)

res = residuals(tsmod1)
res.ts = xts(res[c(-1,-2)], order.by = as.Date(air$Date[c(-1,-2)]))
grid.arrange(ggacf(res.ts), ggacf(res.ts, type="partial"), ncol=2)
log(length(res.ts))
Box.test(res.ts, lag=6, type='Ljung-Box') # It's white noise!
```

```
res.ts
qqnorm(res.ts)
acf(res.ts, lag.max=365)
#########################################################################
# FITTED VISUALIZATIONS

Actual = air.ts$T
Fitted.Reg = xts(mod4$fitted.values, order.by=as.Date(air$Date))
Fitted.ARMA = xts(mod4$fitted.values+tsmod1$fitted.values, order.by=as.Date(air$Date))
Fitted.ts = cbind(Actual, Fitted.Reg, Fitted.ARMA)
plot.xts(Fitted.ts[,c(-3)], legend.loc="topright")
plot.xts(Fitted.ts[,c(-2)], legend.loc="topright")

window(Fitted.ts, start="2004-09-01")
rmse <- function(actual, pred) {
  rmse <- sqrt(t(actual-pred)%*%(actual-pred)*(1/length(actual)))
  rmse
}
rmse(Fitted.ts$T, Fitted.ts$Fitted.Reg)
rmse(Fitted.ts$T[c(-1,-2)], Fitted.ts$Fitted.ARMA[c(-1,-2)])
```

*Summary of **mod4** (best performing linear regression model with 4 predictors)*

```
> summary(mod4)

Call:
lm(formula = T ~ CO + C6H6 + NOx + NO2, data = air.ts)

Residuals:
    Min      1Q  Median      3Q     Max
-10.191  -2.530   0.000   2.411  17.455

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.124923   3.103713   9.062  < 2e-16 ***
CO          -0.034868   0.002330 -14.964  < 2e-16 ***
C6H6        -0.253247   0.090227  -2.807  0.00526 **
NOx         -0.012821   0.001548  -8.285 1.97e-15 ***
NO2          0.028734   0.000920  31.234  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.789 on 386 degrees of freedom
Multiple R-squared:  0.7685,    Adjusted R-squared:  0.7661
F-statistic: 320.4 on 4 and 386 DF,  p-value: < 2.2e-16
```

Results of Shapiro-Wilk normality test on residuals of mod4

```
> shapiro.test(e)

        Shapiro-Wilk normality test

data:  e
W = 0.98293, p-value = 0.0001419
```

Summary of ARMA(1,0) on residuals of mod4

```
Model:
ARMA(1,0)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2155 -1.3793  0.1278  1.4297 13.4471

Coefficient(s):
          Estimate  Std. Error  t value Pr(>|t|)
ar1       0.789869    0.031099   25.398   <2e-16 ***
intercept 0.002072    0.117039    0.018    0.986
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Fit:
sigma^2 estimated as 5.37,  Conditional Sum-of-Squares = 2088.81,  AIC = 1770.79
```

Summary of ARMA(1,2) on residuals of mod4

```
Model:
ARMA(1,2)

Residuals:
     Min       1Q   Median       3Q      Max
-7.41777 -1.34165  0.05303  1.45786 12.99736

Coefficient(s):
          Estimate  Std. Error  t value Pr(>|t|)
ar1       0.816113    0.047460   17.196   <2e-16 ***
ma1       0.021656    0.070536    0.307   0.7588
ma2      -0.106667    0.064257   -1.660   0.0969 .
intercept 0.007841    0.106628    0.074   0.9414
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Fit:
sigma^2 estimated as 5.318,  Conditional Sum-of-Squares = 2063.19,  AIC = 1770.97
```

Summary of ARMA(2,1) on residuals of mod4

```
Model:
ARMA(2,1)

Residuals:
      Min        1Q    Median        3Q       Max
-7.371283 -1.280024 -0.001207  1.457969 13.221711

Coefficient(s):
           Estimate  Std. Error  t value Pr(>|t|)
ar1       -0.123747    0.037876   -3.267  0.00109 **
ar2        0.707673    0.034522   20.499  < 2e-16 ***
ma1        0.971863    0.014547   66.810  < 2e-16 ***
intercept -0.002451    0.224567   -0.011  0.99129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Fit:
sigma^2 estimated as 5.214,  Conditional Sum-of-Squares = 2022.87,  AIC = 1763.25
```

Summary of ARMA(2,2) on residuals of mod4

```
Model:
ARMA(2,2)

Residuals:
     Min       1Q   Median       3Q      Max
-7.41618 -1.32918  0.04526  1.46074 12.98777

Coefficient(s):
           Estimate  Std. Error  t value Pr(>|t|)
ar1        0.749535    0.435944    1.719   0.0856 .
ar2        0.052050    0.337716    0.154   0.8775
ma1        0.087348    0.434087    0.201   0.8405
ma2       -0.103482    0.068012   -1.522   0.1281
intercept  0.007883    0.114628    0.069   0.9452
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Fit:
sigma^2 estimated as 5.317,  Conditional Sum-of-Squares = 2063.07,  AIC = 1772.95
```

Summary of ARMA(1,1) on residuals of mod4

```
Model:
ARMA(1,1)

Residuals:
     Min       1Q   Median       3Q      Max
-7.46576 -1.33302  0.08113  1.45300 13.51057

Coefficient(s):
           Estimate  Std. Error  t value Pr(>|t|)
ar1       0.7582051   0.0442685   17.127   <2e-16 ***
ma1       0.0827161   0.0729279    1.134    0.257
intercept 0.0008349   0.1264925    0.007    0.995
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Fit:
sigma^2 estimated as 5.352,  Conditional Sum-of-Squares = 2082.01,  AIC = 1771.52
```