STAT 443: Prediction of Monthly CO levels in Madrid

Sally Bagk, Naitong Chen, Huaiwen Dong, Xinyi Yan, Chloe You

March 28, 2019

Summary

This report explores different models for predicting monthly average carbon monoxide levels using data collected from the weather Station Moratalaz in central Madrid and forecasts monthly CO levels for the year 2018 using the best model. Out of the eight exponential, ARIMA, ARIMAX and SARIMA models studied, Holt-Winters additive seasonal model is found to be the best in terms of out-of-sample RMSE, followed by the SARIMA model. This is likely due to the strong trend and seasonality patterns in the data. For the same reason, the averaging model produce the largest out-of-sample RMSE values, as it does not address seasonality. Autoregressive and persistence models produce slightly better results, as they only take the previous few observations and the change in CO levels between months are relatively gradual. The exogenous variables in ARIMAX improves the out-of-sample RMSE compared to autoregressive models.

Introduction

Previous studies have shown the effect of long-term CO exposure on human health[1]. This gas is known to cause headache, nausea, vomiting and more. It is even known to be fatal in the extreme or cause long term health concerns[2]. It is not possible to see or smell this gas, but it can be harmful to our bodies. Although often neglected as an important component of air pollutants, we see the essence of being able to forecast CO levels in assisting to address public health issues. This report discusses different methods to predict future months' CO levels by using various time series models.

While persistence and averaging of past values act as sufficient base models for camparison, exponential smoothing methods such as Holt-Winters Additive Seasonal give important insight to the trend and seasonal effects present in the data. Other ARIMA/ARIMAX/SARIMA models that incorporate previous observations, other explanatory variables, and seasonal effects are then explored to find the strongest predictive time series model.

We begin with a description of the dataset used in this study, along with exploratory analysis that provides information on the range of monthly CO levels in Madrid and helps determine candidate forecast models. Then all eight models studied are discussed in terms of goodness of fit along with suggested explanations. Finally, the two best models in terms of out-of-sample RMSE are used to make predictions of average monthly CO level in 2018, followed by a comparison of the forecasted values.

Methods

Description of Data

The raw data sourced from Kaggle[3] consists of hourly concentrations of CO, SO_2, NO_2 from January 2001 to April 2018. To make all patterns of the data obvious, we aggregate by calculating the monthly average concentrations for each of the pollutants. Since the missing rate is quite low for all three variables (1% or less except 5% for March 2006), the missing values are removed when calculating the monthly averages. The monthly CO level is our response variable.

Although the data is available up to April of 2018, since it does not contain the full year, only data from 2001 to 2017 are used, with data from 2001 to 2016 being in the training set and data from 2017 being the holdout set.

Upon checking the correlation of all three variables in the dataset, it is found that NO_2 and SO_2 levels are strongly correlated with the response (0.7127, 0.7859 respectively). Hence, these variables are both included as potential exogenous variables. See below.

Variable	Type	Description
CO	Response	Carbon Monoxide level measured in mg/m^3
NO_2	Exogenous in ARIMAX	Nitrogen dioxide level measured in $\mu g/m^3$
SO_2	Exogenous in ARIMAX	Sulphur dioxide level measured in $\mu g/m^3$

Table 1: Description of all variables

The summary statistics of all three variables is as follows. These values contain information of the magnitude and variability of the monthly average CO, NO_2, and SO_2 levels and provide context when we evaluate how well the forecast models perform.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Count	Std. dev.
CO	0.117	0.273	0.399	0.441	0.525	1.540	198	0.220
NO_2	20.942	39.150	48.029	47.559	56.448	78.713	198	12.567
SO_2	1.084	6.095	8.797	10.512	15.171	31.466	198	6.309

Table 2: Summary Statistic of CO (top, in mg/m^3*), NO_2 (middle, in* $\mu g/m^3$*) and SO_2 (bottom,* $\mu g/m^3$*) at Station Moratalaz in Madrid*

We now look at the time series plot as well as the autocorrelation and partial autocorrelation plots of the response variable from 2001 to 2016 on the next page. It illustrates the overall decreasing trend as well as the strong seasonal effect, which motivates the trend and season analysis, as well as choosing the SARIMA model as a potential forecast model. At the same time, it helps us determine the order of ARIMA models that would fit the data well, which is further discussed in the Model Fitting section.

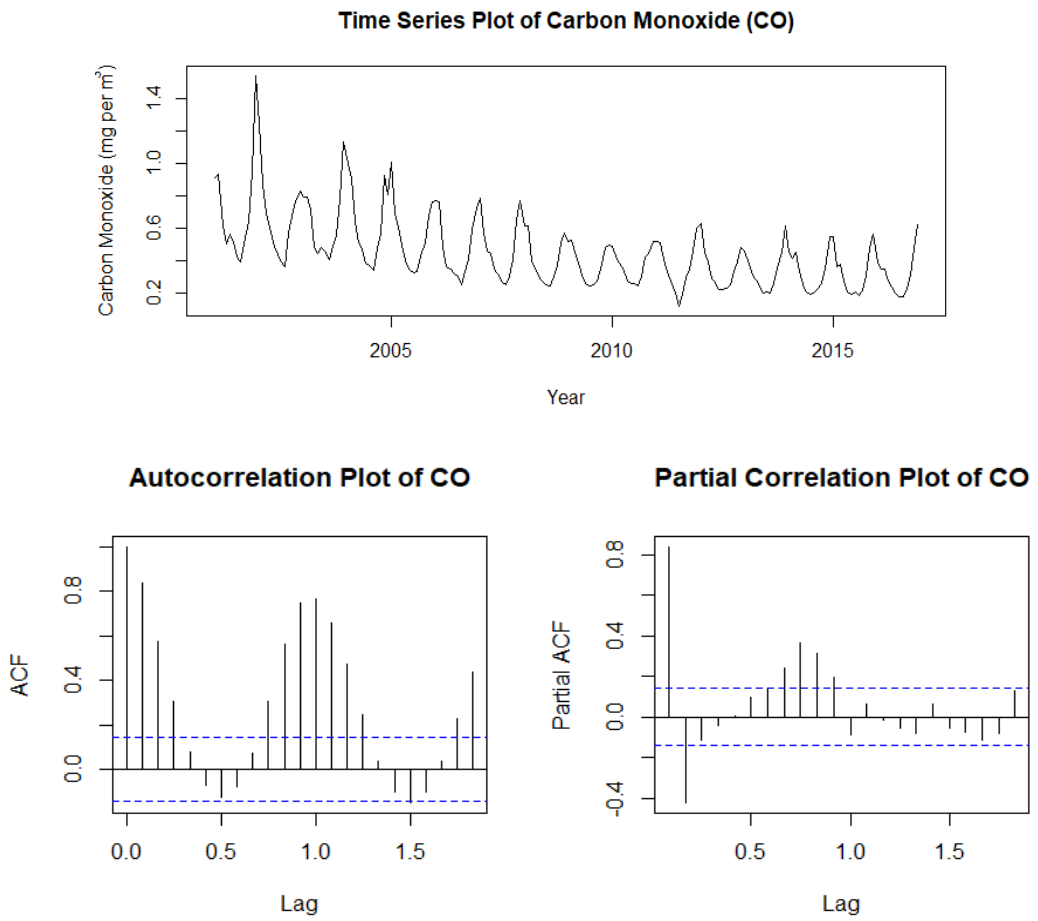

Figure 1: (a) Time series plots (top), (b) ACF plot (bottom left) and (c) PACF plot (bottom right) of monthly average Carbon Monoxide levels at Station Moratalaz in Madrid from 2001 to 2016

Model Fitting

There are a variety of models to be explored. As R provides the in-sample RMSE, this value is first used to check if the model may be a good fit. Given that the in-sample RMSEs do not appear to be too high compared to the sample standard deviation, the out-of-sample RMSEs are calculated for those models and compared. The two models with the lowest out-of-sample RMSEs are used for predicting the carbon monoxide levels in Madrid.

Baseline

Persistence and averaging are typically the first models to be fitted for time series data. This gives a baseline of the in/out-of-sample RMSEs to be compared against. Upon analysis, the out-of-sample RMSEs for both models are 0.079 and 0.180 respectively. We compare these two values with the other models studied and find persistence performs almost as well as some of the autoregressive models, yet averaging is the worst of all models. This is likely due to the strong seasonality pattern and relatively gradual change in average CO levels between every two months. The persistence model takes advantage of the fact that previous observation is correlated with the current one, yet the averaging model completely ignores the seasonality effect.

Holt-Winters with Trend and Additive Seasonality

Holt-Winters additive seasonal exponential smoothing method is found to be the strongest model (lowest out-of-sample RMSE of 0.036). Since there is a clear seasonality and trend observed from the CO time series plot, simple and linear exponential smoothing methods are not the best fit. *Figure 2* below is the autocorrelation of deseasonalized and detrended CO level data that shows the now absent seasonality and trend. Only insignificant white noise is left after fitting the model. A similar approach was taken for Holt-Winters multiplicative model. However, this model is concluded to be an inappropriate fit given the high in-sample and out-of-sample RMSEs.

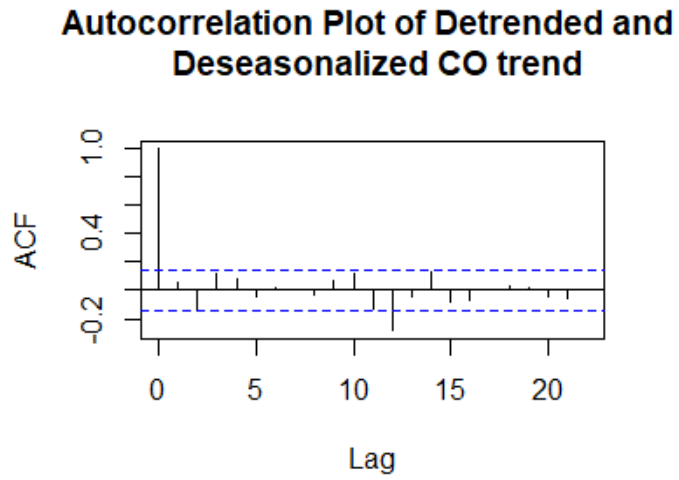

Figure 2: ACF plot of residuals from the Holt-Winters model after detrending and deseasonalizing shows there is mostly only white noise left

Autoregressive Models

In *Figure 1b* and *1c*, we observe damped sine waves in the ACF plot while the PACF plot cuts off at lag 2, indicating that an AR(2) model would be a reasonable choice. The AR(2) model leads to an out-of-sample RMSE of 0.074. Out of interest, AR(1) and AR(3) are also fitted. AR(3) is found to have the same out-of-sample RMSE as AR(2) up to three decimal places and AR(1) has the highest out-of-sample RMSE with 0.075. All three out-of-sample RMSE values are not significantly lower than the baseline persistence out-of-sample RMSE of 0.079, indicating that autoregressive models which regress on previous 1 to 3 observations do not explain the current observation well. This aligns with the suggested explanation that the persistence model fits well, as both the persistence model and the autoregressive models predict the current value using previous observations.

ARIMAX(2,0,0)

Next, we hope to build upon our AR(2) model and use exogenous variables to better explain the CO levels. As mentioned previously, SO_2 and NO_2 are some of the pollutants which are the strongest correlated with CO, the pollutant we are interested in predicting. From the cross-correlation plots in *Figure 3,* we see that with no surprise, at lag 0, these potential exogenous variables seem to have the strongest correlation with average CO levels. However, since the goal of the study is to forecast average monthly CO levels, we choose to use previous lag one to three of NO_2, and previous lag one and two of SO_2, which still have strong correlations, in the ARIMAX(2,0,0) model. We obtain the out-of-sample RMSE as 0.070, which has a relatively small improvement compared to the AR(2) model.

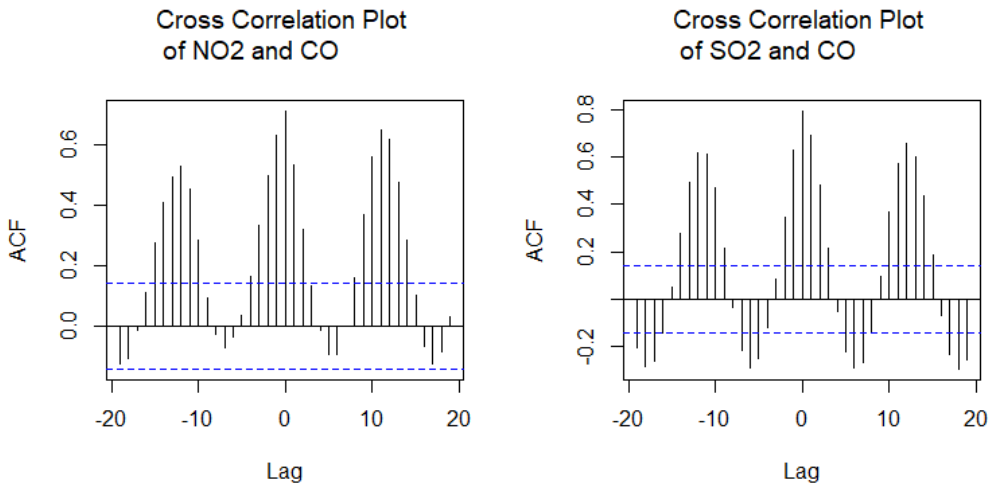

Figure 3: Cross-correlation plot of (a) NO_2 levels (left) and (b) SO_2 levels (right) against CO at Station Moratalaz in Madrid to determine best lags to use for predicting monthly average CO levels

SARIMA(2,0,0)(0,1,0)$_{12}$

The data contains obvious seasonality with period length equal to 12. Upon observing the ACF and PACF plots by month, a SARIMA model with order 1 seasonal difference is fitted, giving an out-of-sample RMSE of 0.062. To further show that this model is a good fit, we checked the autocorrelation of the model residuals (*Figure 4* below). Although we observe one spike at a high lag order, overall, the ACF plot seems like white noise and hence indicates an appropriate choice of model. As stated above, we expect a good fit with this model, because it addresses the strong seasonal pattern in the data.

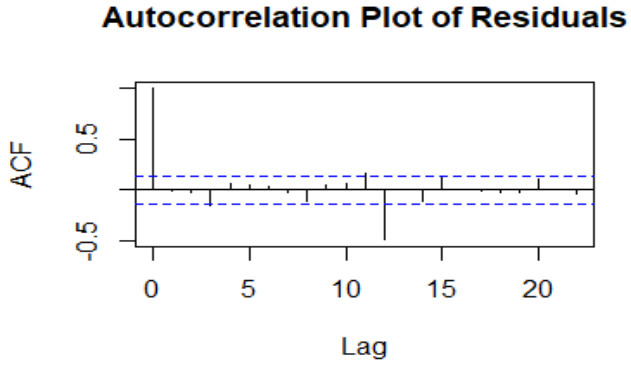

Figure 4: ACF plot of residuals to check for only white noise remaining after fitting a SARIMA model

Forcasting 2018 Carbon Monoxide Level

As Holt-Winters and SARIMA(2,0,0)(0,1,0)$_{12}$ have the lowest out-of-sample RMSE, these models are used to forecast for 2018. We retrain the models, this time with all data from 2001 to 2017, and produce *Figure 5* below, which shows that the forecast of 2018 CO level in Madrid. We notice the forecasts maintain the overall decreasing trend of CO levels relative to the past years, which is consistent with what we observe from 2001 to 2016. Notice that the Holt-Winters model produces forecast values slightly lower than those of the previous year, but SARIMA's predictions remain on the same level. This may correspond to the cut of gas emission promoted by the Spanish government as a

way of clearing the atmosphere[4]. The forecast also follows the annual trend of having a lower CO level around the middle of each year. The exact values of the 2018 forecast can be found in *Table 5* in the appendix.

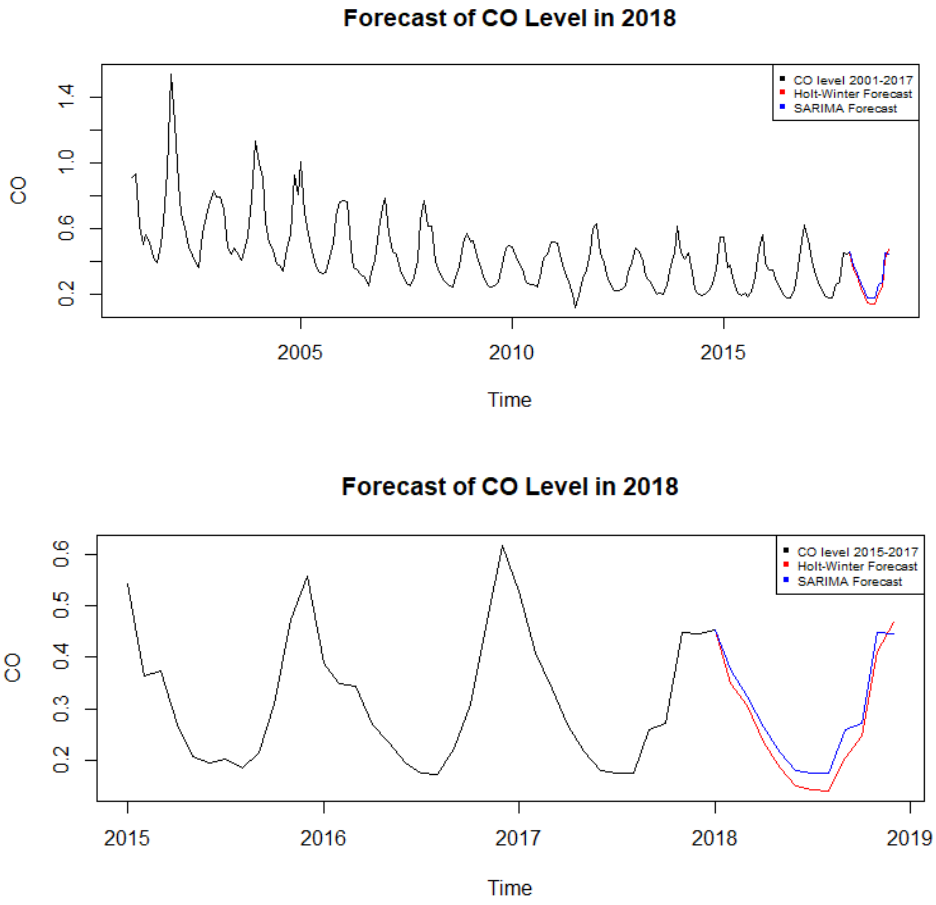

Figure 5: (a) 2018 carbon monoxide forecasts by Holt-Winters and SARIMA models (top) and (b) zooming in on the last five years (bottom)

Results

Table 3 summarizes the out-of-sample RMSEs for all models that are discussed in this report. Since Holt-Winters and SARIMA are the best models, *Table 4* summarizes predicted values for 2017 against the actual for the two models as well as the baseline persistence, averaging and AR(2) models. Although Holt-Winters does not always provide the closest prediction value to the holdout set values (i.e. April), overall Holt-Winters results in the closest predictions.

Persistence	Averaging	Holt-Winters	AR(1)	AR(2)	AR(3)	ARIMAX(2,0,0)	SARIMA
0.079	0.180	0.036	0.075	0.074	0.074	0.070	0.062

Table 3: Summary of out-of-sample RMSEs of all models discussed

2017 Months	Holdout	Persistence	Averaging	AR(2)	Holt-Winters	SARIMA
January	0.528	0.618	0.450	0.645	0.435	0.414
February	0.409	0.528	0.450	0.469	0.374	0.409
March	0.339	0.409	0.450	0.363	0.377	0.370
April	0.269	0.339	0.449	0.331	0.288	0.267
May	0.217	0.269	0.448	0.277	0.234	0.233
June	0.179	0.217	0.447	0.245	0.198	0.187
July	0.175	0.179	0.446	0.222	0.183	0.168
August	0.175	0.175	0.444	0.234	0.177	0.173
September	0.259	0.175	0.443	0.235	0.218	0.222
October	0.270	0.259	0.442	0.339	0.306	0.325
November	0.448	0.270	0.441	0.315	0.446	0.453
December	0.445	0.448	0.441	0.526	0.444	0.609

Table 4: Comparison between actual and forecasted average monthly CO levels in 2017 of selected discussed models

Conclusion

Holt-Winters is the best model to predict the carbon monoxide levels followed by SARIMA$(2,0,0)(0,1,0)_{12}$. These two models are the only models which explain the seasonality effect, hence are used to forecast the 2018 CO levels. Averaging is the worst model. This is as expected as it does not explain the strong seasonality effect present in the data. ARIMAX model is better than ARIMA models with the orders picked from visualizing the ACF and PACF plots, due to the strong correlations between the exogenous variables at the selected lags and the response.

Improvements and Next Steps

Doing further research on potential effects on carbon monoxide levels can help to strengthen the predicting ability of the ARIMAX model. However, taking seasonality into account, SARIMAX model may be best for this data.

It is also possible that other SARIMA models where p and q orders at the seasonality level are greater than zero can help to predict carbon monoxide levels. However, these models are not discussed in STAT443.

Contributions

Our team consists of a close group of friends who are all statistics majors and eager to apply times series knowledge to a practical dataset. The author names are listed alphabetically by surname. Sally contributed mostly to the writing, statistical analysis and constructive criticisms. Naitong contributed mostly to exploratory analysis and recommended models to fit. Huaiwen and Xinyi verified validity of the code and fixed critical issues. Huaiwen and Xinyi also compiled the data tables and made time series plots and forecasted using the best fit model. Chloe contributed mostly to the coding of the out-of-sample RMSE, initial data cleaning. Overall, everyone contributed equally to this project.

References

1. Sanhueza, Pedro, et al. "Health risk estimation due to carbon monoxide pollution at different spatial levels in Santiago, Chile." Environmental monitoring and assessment 167.1-4 (2010): 165-173.
2. Ernst, Armin, and Joseph D. Zibrak. "Carbon monoxide poisoning." New England journal of medicine 339.22 (1998): 1603-1608.
3. Decide Soluciones. Air Quality in Madrid (2001-2018): Different pollution levels in Madrid from 2001 to 2018. (Version 5). Madrid, Spain: Decide Soluciones, 2018. Web. 27 Mar 2019. < https://www.kaggle.com/decide-soluciones/air-quality-madrid/version/5?>
4. Amigo, Ignacio. “Madrid's Getting Close to ‘Plan A’ Rollout to Fight Air Pollution.” *Next City*, 16 Oct. 2017, nextcity.org/daily/entry/madrid-plan-fight-air-pollution.

Appendix

Changing size of training and holdout

If the training set is changed from 16 years to 15, in order to have 2 years or 24 months of data in the holdout set, the out-of-sample RMSEs are summarized in Table 5. The conclusion remains similar in that Holt-Winters model has the lowest out-of-sample RMSE followed by (closely) the SARIMA model. However, the AR(2) model has a lower out-of-sample RMSE than ARIMAX for the 2-year holdout set.

Persistence	Average	Holt-Winters	AR(1)	AR(2)	AR(3)	ARIMAX(2,0,0)	SARIMA
0.085	0.186	0.056	0.083	0.076	0.079	0.077	0.058

Table 5: Summary of out-of-sample RMSEs of all models with a 2-year holdout set

2018 forecasted values

The forecasted monthly carbon dioxide levels for 2018 by Holt-Winters and SARIMA$(2,0,0)(0,1,0)_{12}$ can be found in *Table 8*.

2018 Months	Holt-Winters	SARIMA
January	0.452	0.453
February	0.348	0.377
March	0.307	0.325
April	0.234	0.263
May	0.183	0.214
June	0.148	0.178
July	0.142	0.175
August	0.140	0.174
September	0.204	0.259
October	0.246	0.270
November	0.410	0.448
December	0.469	0.445

Table 6: 2018 forecasted values