Univariate Time Series Prediction of CO Contaminants in an Italian City

Jing Hao

Background and Research Questions

Carbon monoxide is a gas and is found in air. High levels of carbon monoxide are poisonous to humans and, unfortunately, it cannot be detected by humans as it has no taste or smell and cannot be seen. Making accurate predictions of concentration of air pollution chemicals in any environment is complex considering the physical and chemical properties associated with many factors. Time series is one of the methods that can be used to predict air contaminates.

In this analysis, I will use univariate time series models ARMA/ARIMA to predict CO contaminates concentration levels in an Italian city.

Data

The data is obtained from UCI Machine Learning Repository. It was recorded by 5 metal oxide chemical sensors located in a significantly polluted area in an Italian city, and I will analyze one of them, CO. The dataset contains 9358 instances of hourly averaged responses spreading from March 2004 to February 2005.
Below is the attributes description loaded from UCI Machine Learning Repository.

0 Date (DD/MM/YYYY)
1 Time (HH.MM.SS)
2 True hourly averaged concentration CO in mg/m^3 (reference analyzer)
3 PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
4 True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
5 True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
6 PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
7 True hourly averaged NOx concentration in ppb (reference analyzer)
8 PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
9 True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
10 PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
11 PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
12 Temperature in Â°C
13 Relative Humidity (%)
14 AH Absolute Humidity

Data Wrangling

Missing values and inappropriate data type

I first dropped the columns (“Unnamed: 15”, “Unnamed: 16”) that have empty values. I then dropped rows with NaN values. This results 9357 records.

Within these observations, there are missing values marked as “-200”, which were covered to “NaN” to be processed later.

By checking the data type of each variable, I found that in several numeric variables, the dot (“.”) in float numbers were mismarked as comma (“,”). I cleaned it and converted them to float numbers. With more “-200” values found in this process, I have replaced it to “NaN” as well.

With target variables identified as S1, S2, S3, S4, and S5, I replaced the NaN values with the column mean.

Time Series Conversion

The default index of the dataframe is 0, 1, 2, 3, …. To obtain a datetime index, I combined the “Date” and “Time” columns as “Datetime”, converted its data type from string to datetime (stored in “DateTime” , and replaced the default index with “DateTime” .

A dataframe consists of S1-S5 columns and DateTime index was generated for further analysis. (S2, S3, S4, and S5 will be saved for further multivariate analysis in a separate project.)

Exploratory Data Analysis

The following is time series plots of S1 (CO) concentration level.

Some features of the plots:

1. There is no consistent trend (upward or downward) over the time.
2. It shows seasonality of daily repeated patterns for all five variables. For example, when I looked at the plots for a shorter period, from '2004-10-04' to '2004-10-07', similar patterns repeated day after day within each variable.
3. No obvious outliers have been observed.
4. The variability is nearly constant over the time.

I also ran first-order time series lag plots for each target variables in order to check whether the time series is random or not. The plots suggest a moderate correlations exist for each variable.

Check Stationarity

A time series is said to be weakly stationary if it has:

1. Constant mean over the time;
2. Constant variance over the time;
3. And its autocorrelation does not depend on the time.

Stationarity of a time series is important because most time series models are based on the assumption that the series is weakly stationary. In this project, we use the following methods to check the stationarity:

1. Plotting Rolling Statistics - by plotting the moving average or moving variance and see if it varies with time.
2. Dickey-Fuller Test - a statistical test used for checking the stationarity giving the assumption that the time series is not stationary.

With nearly constant rolling mean and rolling variance, and with p-value of Dickey-Fuller Test close to zero, I can conclude that S1is weakly stationary.

Seasonal ARIMA

ARIMA model is a popular regression method used for weakly stationary univariate time series. It is a combination of autoregressive, differencing terms, and moving average model. Autoregressive models are based on the idea that the current value of series can be explained as a function of past p values. A moving average term in a time series model is a past error (multiplied by a coefficient). the moving average model of order q , abbreviated as MA(q), assumes the white noise wt on the right-hand side are combined linearly to form the observed data.

A seasonality has also been observed with S=24 (hours per day). Almost by definition, it may be necessary to examine differenced data when we have seasonality. Seasonality usually causes the series to be nonstationary because the average values at some particular times within the seasonal span (months, for example) may be different than the average values at other times.

As a summary, I will use model $\mathrm{ARIMA}(p, d, q) \times (P, D, Q)S$.
With p = non-seasonal AR order,d = non-seasonal differencing, q = non-seasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = time span of repeating seasonal pattern. These parameters will be determined based on the patterns of ACF and PACF plots.

The ACF(Autocorrelation Function) is defined as correlations between x_t and x_{t-h} for h = 1, 2, 3, etc. It measures the linear predictability of the series at time t, using only the values x_{t-h} .

PACF (Partial Autocorrelation Function) is a conditional correlation under the assumption that we know and take into account the values of some other set of variables. For a time series, the partial autocorrelation between x_t and x_{t-h} is defined as the conditional correlation between x_t and x_{t-h}, conditional on x_{t-h+1}, ... , x_{t-1}, the set of observations that come between the time points t and t−h.

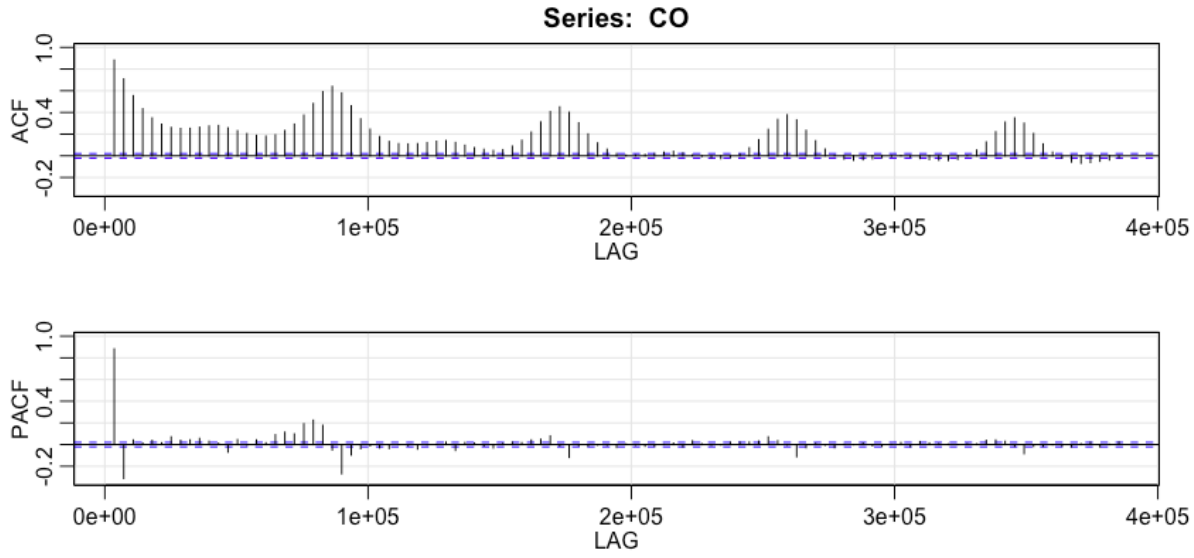
Results

ACF and PACF

Below ACF and PACF plots for S1 have shown seasonal patterns. The patterns repeat with every 24 observations, so that we can confirm S=24. The lag values on x axis are relatively large with the measuring unit in seconds.

For the seasonal part, ACF tails off, and PACF cuts off after the first or second repeats.

For the nonseasonal part, ACF tails off, and PACF cuts at lag = 2.



Model Tuning and Results

With several sets of parameters tested, (2,0,0,1,1,1,24) gives the best results with the lowest AICc and BIC.

Parameters (p, d, q, P, D, Q, S)	AICc	BIC
2,0,0,1,0,0,24	9.921957	8.924797
2,0,0,1,0,1,24	9.750293	8.753896
2,1,0,1,0,1,24	9.796905	8.800508
1,0,0,1,0,1,24	9.758126	8.760965
1,0,0,0,1,1,24	9.760428	8.762504
1,0,0,1,1,1,24	9.745689	8.748529
2,0,0,1,1,1,24	**9.739171**	**8.742774**
2,0,0,1,1,0,24	10.04157	9.044411

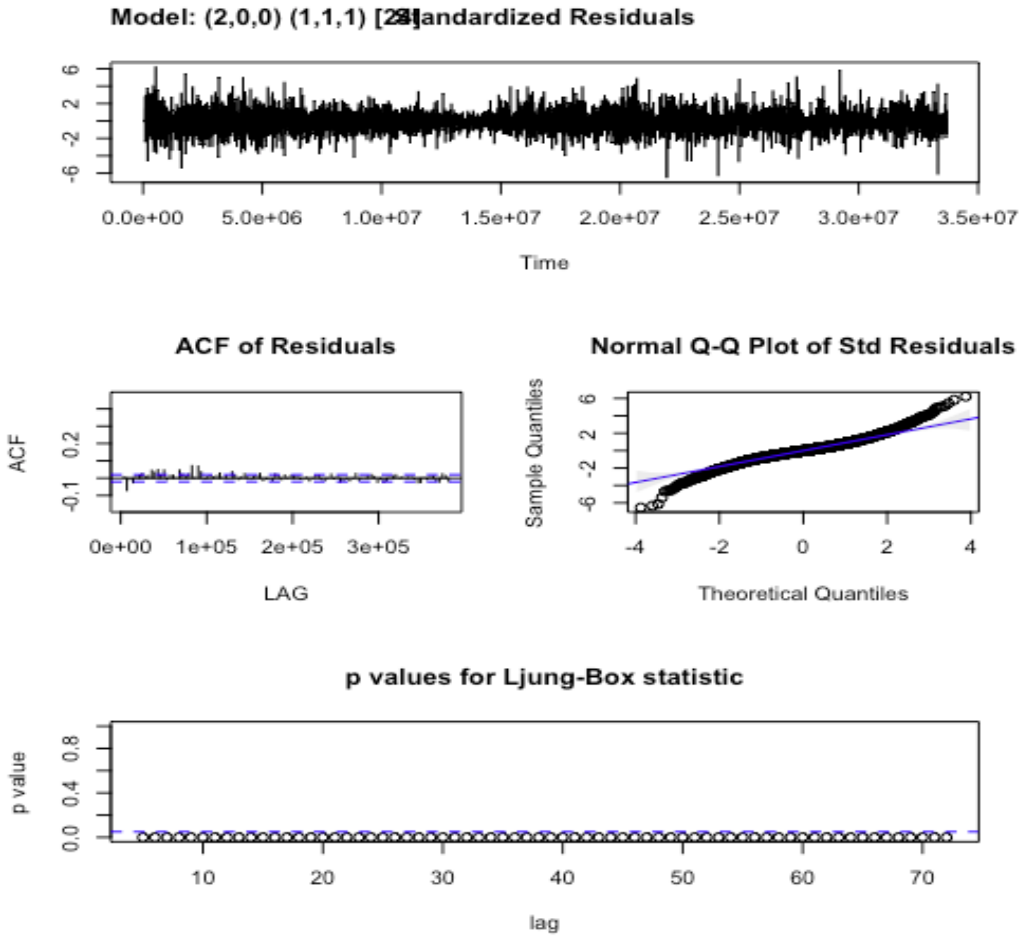
The t-table below gives estimates for all parameters. P-values for ar1, ar2, sar1, sma1 are close to zero showing a strong evidence against the null.

$ttable

	Estimate	SE	t.value	p.value
ar1	0.9459	0.0104	90.8291	0.0000
ar2	-0.0819	0.0103	-7.9199	0.0000
sar1	0.1319	0.0116	11.3469	0.0000
sma1	-0.9375	0.0046	-203.4361	0.0000
constant	-0.0136	0.0188	-0.7240	0.4691

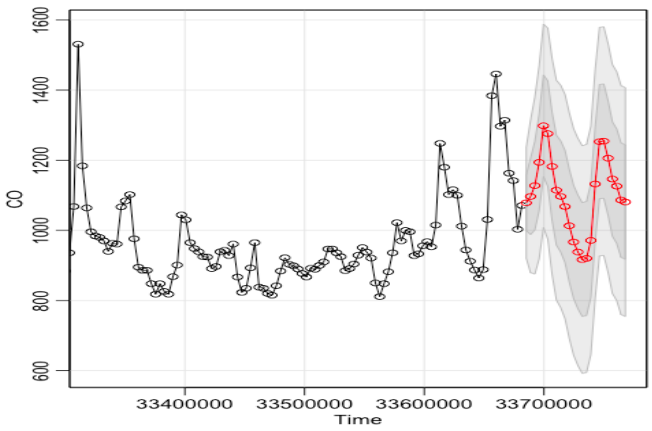
ar1 - value this hour related to previous hour x_{t-1}
ar2 - value this hour related to x_{t-2}
sar1 - value this hour related to same time yesterday.
sma1 - value this hour related to the noise same time yesterday.

The residual plots look fine too.



Forecast

Below plots show the 24 forecasts using model we have discussed. The dark grey area indicates 95% CI, and the light grey areas indicates 90% CI.



Conclusion and Further

From this analysis, we can conclude that the CO concentration in the air can be predicted using the values within the past two hours together with the values at the same time yesterday. Based on this, we may further analysis the peak values during the hours of day and investigate and control any factors/human activities causing this. Since the data also provided other four contaminates, in my next project I will do a multivariate time series analysis.

Reference

S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical, Volume 129, Issue 2, 22 February 2008, Pages 750-757, ISSN 0925-4005.

R. Shamway, D. Stoffer, Time Series Analysis and Its Applications. Third Edition. Page 83-154.