Interpret and predict concentration for carbon oxide with several statistical models in an urban pollution monitoring scenario

Long Wan

Abstract

This paper interpreted the actual fluctuation of concentration for carbon oxide with measured concentrations for several kinds of gas by an air quality multisensory device and other exogenous variables, and then made predictions for actual CO concentration with several time-series models, MA, ARMA and VAR. Model outputs were compared. Results obtained include: 1) device measured concentration for CO, NMHC, NOx, NO2 and relative humidity have strong linear relationships with and Granger-cause to actual concentration for CO; 2) ARIMA model reflects more involvement of noise than MA and VAR model on predicting; 3) VAR, ARIMA and MA model are better than linear model on predicting CO concentration. Finally, some analysis on whether to apply these model are offered.

Keywords: time series; CO concentration; multisensory device; air quality; on-field calibration

1. Introduction

With the development of the economy and improvement of people's living standard, air pollution has become a serious issue in almost all countries in the world. There are 5.5 million people died of diseases induced by air pollution every year, including chronic bronchitis, bronchial asthma and lung cancer. Studying and monitoring pollutant composition is an extremely essential subject in air pollution alleviation and human health.

As a post-industrialized country and the country on wheels, the United States is also facing air pollution issue, mainly because of its high auto possessions. In 2015, the total number of vehicles are 285 million in US, while auto per capita ranked 1st around the world, placing vehicle emission one of the largest pollutant source all over the country. American people are accustomed to driving cars when leaving for working and shopping and public transportations are underdeveloped and underused.

Car emissions are dangerous. Its composition mainly include carbon oxide(CO), nitric oxide(NOx), nitric dioxide(NO2), non-metallic hydrocarbons(NMHC) and benzene(C6H6), which are all lethal substances for human beings. Meanwhile, through chemical reaction under the sunshine, nitric oxide and benzene might generate another substance, ozone(O3), which is harmful to lung and might cause death if breathed too much.

From macro level, governments are able to reduce the average harm of air pollution by policy enforcements, information disclosure and resource allocation. But variance might exist at individual level. Individuals should pay attention to air pollution and take measures by ourselves to suppress air pollution and avoid harms elicited by it. Existing air quality sensor devices were expensive for personal use, indeed. And there were few devices that could detect multiple kinds of pollutant concurrently. Plus, air quality information release was not frequent and always hysteretic. So people could hardly purchase an air quality detection by themselves and were impossible to get access to instant pollutant concentrations so that we were often in the dark and no protections were taken though serious air polluting issue occurred.

Scientists developed a new multisensory device able to detect concentrations for CO, NMHC, NOx, NO2 and O3 simultaneously, aiming at monitoring auto emissions instantly for personal use. However, sensor responses toward gas concentrations are not always accurate since detecting 5 kinds of gas at the same time bring some noises to estimation. So before putting it into industry use, they had to calibrate electronic noses, equipped with an additional reliable calibration able to counter specificity and stability issues of solid-state sensors they rely on. In order to achieve this goal, a test should be arranged to see the difference between actual values and detected values. This device was placed on field in a significantly polluted area, at road level, within an Italian city. Besides, a co-located reference certified analyzer was also used for recording the ground truth averaged overall gas concentration for different kinds of gases. The test lasted from March 2004 to February 2005 representing the longest freely available recordings on field deployed air quality chemical sensor devices responses. As CO accounts for the largest portion of emission pollutants, I mainly focused on concentration for CO. Basing on the data detected by the sensor and actual data detected by reference certified analyzer, we could calibrate the

electronic nose, gaining the more accurate estimation of CO concentrations, and predict the trend of it within a short time period. By doing so, a more affordable daily air quality sensor device might be available for ordinary people, providing accurate and instant detection and prediction service of concentration for polluted air, so that people could take measures ahead of time to reduce the potential risks of exposure to polluted environment.

In 2007, S.De Vito, E.Massera and others developed a neural calibration for the prediction of benzene concentrations. This kind of algorithm was capable to limit the absolute prediction error for more than 6^{th} month, after which seasonal influences on prediction capabilities at low-concentrations suggested the need for a further calibration.

In 2009, S.De Vito, Marco Piga and others presented a multivariate calibration for CO, NO2 and NOx with the use of two weeks long on-field data recording and neural regression systems. But for these pollutants, no significant performance boost was detectable when longer recordings were used. Also, there were some trade-off between stability and reliability.

The following parts of the paper is organized as follows. Section 2 describes the data set and presents preprocessing and transformations I have made. Section 3 explains mathematical and statistical theories for several models I chose. Section 4 estimated each model and analyzes the estimation results. Section 5 gives some analysis for applying models. Section 6 concludes the paper and provide further suggestions.

2. Data description and transformation

2.1. General description

The data set can be downloaded from Machine Learning Repository of University of California, Irvine. The original data set contains 15 variables, 4 categories. One are time and date. One consists of 5 ground truth gas concentrations detected by reference analyzer. One contains 5 gas

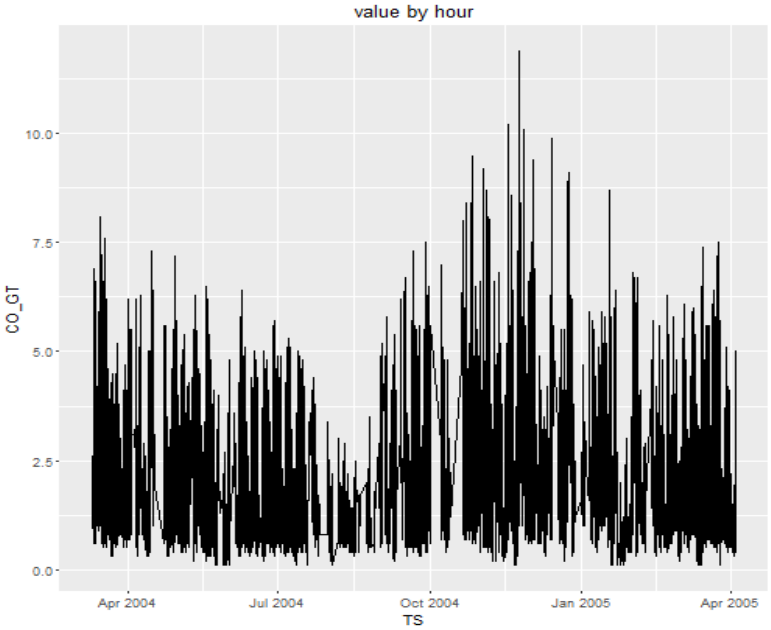

Figure 1 Ground truth for CO concentration, March 2004 - April 2005

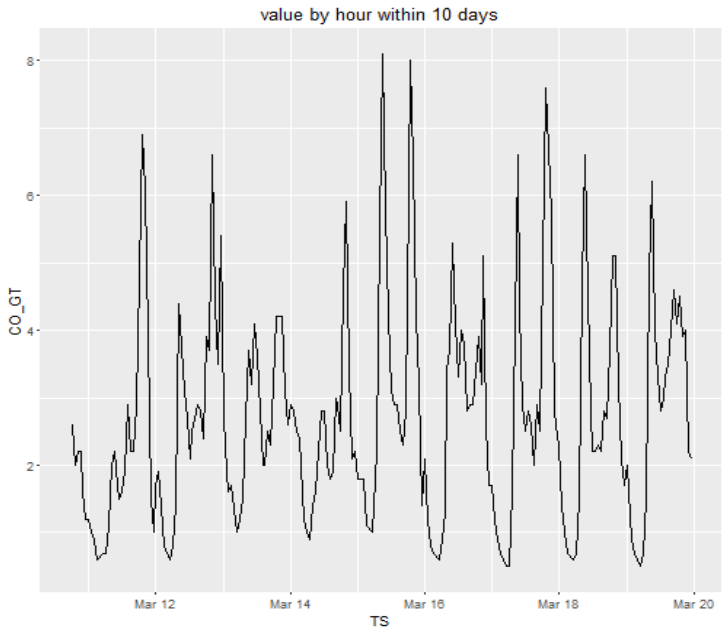

Figure 2 Ground truth CO concentrations of 10 days

concentrations by the targeted multi sensor device. One are some peripheral factors, including temperature, relative humidity and absolute humidity. The latter 3 categories are all numeric. As I only focus on calibration for actual CO concentration, I deleted other 4 ground truth gas concentrations so that there were 11 variables left.

They are hourly data, dating from March 10th, 2004 and ending on April 4th, 2005. Fig.1 shows the line chart of ground truth CO concentration. Since the frequency of collection is pretty high, we could not clearly see the change of CO concentration. But there might be some seasonal changes, CO concentration higher in winter and lower in summer and moderate in spring and fall.

Fig.2 randomly picks up data of 10 days so that features could be explained easily. There are two peaks almost every day, representing two rush hours. Concentrations are always lower during night and higher during daytime, indicating that human activities are highly related to CO emission. It corresponds to our normal understanding.

2.2. Missing values analysis

There are some missing values in the data set. Missing values should be paid attention to since results would be biased when values are not missing randomly. Tab.1 shows the number of missing values for each variable. As we can see, the ground truth CO concentration(CO_GT) has 1683 missing values, while others have 366, which were missing at exactly the same time.

Var	Date	Time	CO_GT	S1_CO	S2_NMHC	S3_NOx	S4_NO2	S5_o3	T	RH	Ah
Num	0	0	1683	366	366	366	366	366	366	366	366

Table 1 Number of missing values for each variable

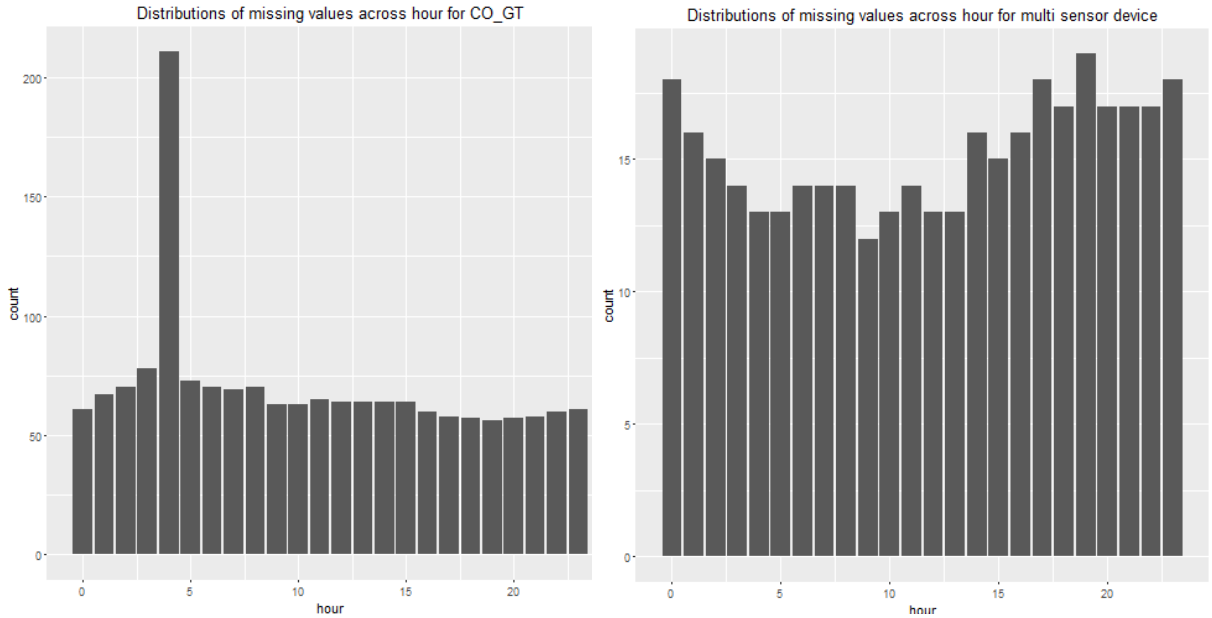

Figure 4 Distribution of missing values across hour.(Left: CO_GT, right: multi sensor device)

Fig.3 shows the distribution of missing values across hours. For CO_GT, missing values distributes almost evenly in each hour except at 4 am. The reference analyzer seemed not work at 4 am in most cases. For the multi sensor device, missing values distributes evenly in each hour. From the hour respective, missing values are quite random and I will dig into 4 am further next.

Fig.4 shows the distribution of missing values across date. There are many dates with only one missing value for CO_GT, most of which are at 4 am. They randomly distributed across the dates. Plus, the analyzer and device might not work for the whole day.

Figure 3 Distribution of missing values across date

Based on previous analysis, most missing values are concentrated on several specific days randomly distributed across the year but not concentrated. Certified reference analyzer did not work at 4 am in most cases, but the dates when it did not work also randomly distributed across the year. So it is safe to say that they are missing completely at random.

The simplest way to handle random missing values is to delete them. After simple deletion, there were 9357 observations left, which was enough to be modeled.

2.3. Transformation

Hourly data is not suitable for following time series modeling. First of all, there are too many consecutive missing values, sometimes even hitting 50. Since breaks are too long in some cases, autoregressive models might be invalid. Secondly, external factors, like rush hours, might influence short-term CO concentrations. We could hardly interpret current CO concentration with that of 1 hour ago.

One way to solve issues above is to turn hourly data into daily data by taking the average value. By doing so, influence of peak hours and long detection breaks would be eliminated. After the transformation, the trend of CO concentration could be more obvious, just as fig.5 shows. In the next few sections, transformed daily data will be used instead of hourly data in estimating and predicting CO concentration. Also, other hourly variables have been taken average.

Another problem is the distribution of variables. The basic assumption of linear model is that variables follow normal distribution. Fig.6 shows density, correlation for each variable and relationships between two variables. As we can see from charts on diagonals, some seem not to follow normal distribution and need to be transformed. Although I didn't know which variables would be chosen in the linear model, square root was taken for CO_GT, S5_O3, and logarithm

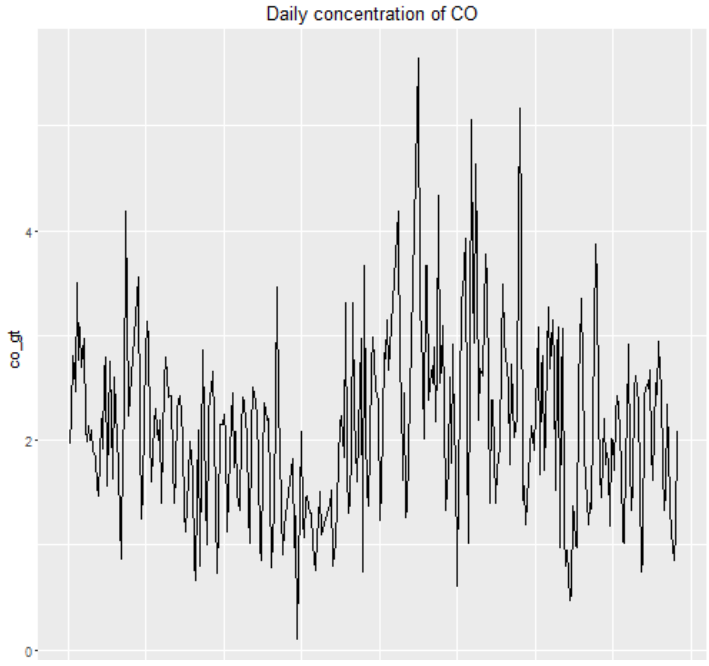

Figure 5 Daily concentration for CO

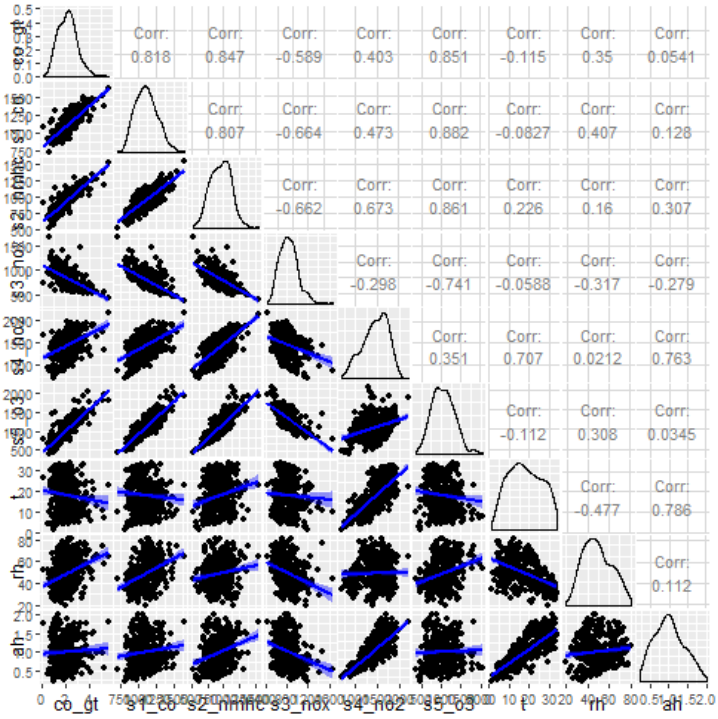

Figure 6 Density, correlation for each variable and linear relationships between two variables

was taken for S1_CO and S2_NOx, so that they would be more likely to follow normal distribution. The linear regression will use both kinds of data, normal transformed and original, and compare the results of linear regression to decide whether to use in future models.

Additionally, from Fig.6, we could find that concentration for CO has positive relationships with S1_CO, S2_NMHC, S4_NO2, S5_O3 and relative humidity, and has negative relationship with S3_NOx. Correlations between CO concentration and absolute humidity, temperature are relatively low compared to with other variables.

3. Models

3.1. Multiple linear model

Given a data set with y_i as the dependent variable and $x_{i,1}, x_{i,2}, \ldots \ldots, x_{i,n}$ as n independent variables, we assume that the relationship between dependent variable and independent variables is linear, and this kind of relationship can be modeled in the following form:

$$y_i = c + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + \alpha_3 x_{i,3} + \cdots + \alpha_n x_{i,n} + \varepsilon$$

where $\alpha_1, \alpha_2, \alpha_3 \ldots \alpha_n$ denote coefficients for each independent variables, c is the constant and ε denotes errors. In this paper, y_i is the CO concentration. Independent variables will be selected based on Akaike information criterior(AIC).

3.2. Autoregressive model

Autoregressive model(AR model) is a representation of a type of random process and time series models. A one-time shock affects the value of the time series data infinitely far into the future. The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term

The notation AR(p) indicates an autoregressive model of order p, and it is defined as:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

where $\varphi_1, \varphi_2, \ldots, \varphi_n$ are the parameters of the model, c is the constant and ε_t is the white noise. In order to maintain the stationarity, parameters should be less than 1. If one parameter was greater than 1, the value of dependent variable will become infinite.

3.3. Autoregressive-moving-average model

Autoregressive-moving-average model(ARMA) model contains two polynomials, one for autoregression as mentioned above and one for moving average, better describing a weakly stationary stochastic process. The AR part involves regressing the variable on its own lagged values just like section 3.2 mentioned. MA part modeling the error term as a linear combination of error term occurring contemporaneously and at various times in the pass.

There are two parameters for ARMA model, p and q. P is the order of autoregressive part and q is the order of moving average part. It can be modeled as follow:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

where c is the constant, ε denotes error terms, φ_i denotes parameters for AR part and θ_i denotes parameters for MA part. Brockwell & Davis recommend using AIC for finding p and q, which will be used in estimation.

3.4. Vector autoregression model

Vector autoregression model(VAC model) is used to capture the linear inter dependencies among multiple time series. Compared to AR model, it allows more than one evolving variable. There are two kinds of VAR, structural VAR and reduced VAR. In this paper, I will use the reduced VAR only, and it can be written in the following form:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t,$$

where y denotes a set of k variables at different time, A denotes coefficients for variable set at each time, and e_t is the error term. The only parameter for VAR model is p, which is the lag order of VAR. The model can be divided into k separate equations for each variable, and my following analysis only take the equation where co_gt is the dependent variable.

4. Estimating and predicting CO concentration

4.1. Unit root test

Before modeling with linear regression and other time series models, data series should be checked stationary. Augmented Dickey-Fuller test was taken for 9 variables. The null hypothesis is that data series is not stationary. Table 2 shows the p-values of each test.

Variable	co_gt	s1_co	s2_nmhc	s3_nox	s4_no2	s5_o3	t	rh	ah
P-value	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.387	<0.01	0.042

Table 2 ADF test results for each variable

All variables but temperature reject null hypothesis at 5% level. So temperature is not stationary and others are considered stationary. In next parts, temperature will be considered in modeling since stationarity is a requirement for linear and time series model.

4.2. Linear regression

Linear regression is estimated for both normal transformed and non-transformed data as mentioned in section 2.3. Variables were chosen based on AIC. The two results are shown in Table 3 and Table 4.

```
Call:
lm(formula = co_gt ~ s1_co + s2_nmhc + s3_nox + s4_no2 + rh,
    data = data3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93494 -0.02474  0.01980  0.05700  0.41244

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.114e+00  8.362e-01  -6.116 2.63e-09 ***
s1_co        5.381e-01  1.042e-01   5.162 4.16e-07 ***
s2_nmhc      1.682e-03  1.018e-04  16.527  < 2e-16 ***
s3_nox       2.158e-01  5.185e-02   4.162 4.00e-05 ***
s4_no2      -2.954e-04  3.496e-05  -8.449 8.62e-16 ***
rh           3.245e-03  6.148e-04   5.279 2.32e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1295 on 341 degrees of freedom
Multiple R-squared:  0.804,     Adjusted R-squared:  0.8012
F-statistic: 279.8 on 5 and 341 DF,  p-value: < 2.2e-16
```

Table 3 Linear regression results for normal transformed data

```
Call:
lm(formula = co_gt ~ s1_co + s2_nmhc + s3_nox + s4_no2 + rh,
    data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.17833 -0.11077  0.03931  0.14505  1.57864

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.7791547  0.2936584 -12.869  < 2e-16 ***
s1_co        0.0014392  0.0002453   5.867 1.05e-08 ***
s2_nmhc      0.0047478  0.0002708  17.534  < 2e-16 ***
s3_nox       0.0006583  0.0001538   4.279 2.44e-05 ***
s4_no2      -0.0008372  0.0000930  -9.002  < 2e-16 ***
rh           0.0094177  0.0016403   5.742 2.08e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3483 on 341 degrees of freedom
Multiple R-squared:  0.8282,    Adjusted R-squared:  0.8256
F-statistic: 328.7 on 5 and 341 DF,  p-value: < 2.2e-16
```

Table 4 Linear regression results for original data

Final model for normal transformed data is:

$$\widehat{\text{co\_gt}}(\text{t}) = -5.11 + 0.54s1\_co + 0.0017s2\_nmhc + 0.22s3\_nox - 0.0003s4\_no2 + 0.0032rh + \varepsilon$$

Final model for original data is:

$$\widehat{\text{co\_gt}}(\text{t}) = -3.78 + 0.0014sqrt(s1\_co) + 0.0047\log(s2\_nmhc) + 0.0007\log(s3\_nox) - 0.0008sqrt(s4\_no2) + 0.0094rh + \varepsilon$$

Two results meet our expectation. Concentration for CO does have positive relationships with factors except NO2, which is not quite same as what was discovered by section 2.3. Mean values of all variables are significant at 0.001% level, showing that average concentration for CO have strong relationships with the average values of 5 variables picked. Results for s1_co and s2_nmhc are in line with expectations. More auto emission, more NMHC, NOx and CO detected at the same time. The higher the relative humidity is, the harder for pollutant to delude. But it is difficult to explain the results for NOx and NO2, since they are different from the description in section 2.3 and do not correspond to the fact that they are emitted simultaneously. Maybe the sensor was not accurate and is not able to reflect the truth, and that comes the calibration. O3 was dropped by AIC selection, maybe it is because O3 does not have any direct relationships with CO_GT since an auto does not emit O3. It is just the derivative of NOx and C6H6.

The second result has a slightly higher R-squared, indicating that combination of variables from original data explained more variance. Plus, p-values are smaller in the second model for each variable. Therefore, even though two models performed similarly, if we have to choose one, the second one is slightly better than the first one. Data will remain unchanged in following time series models.

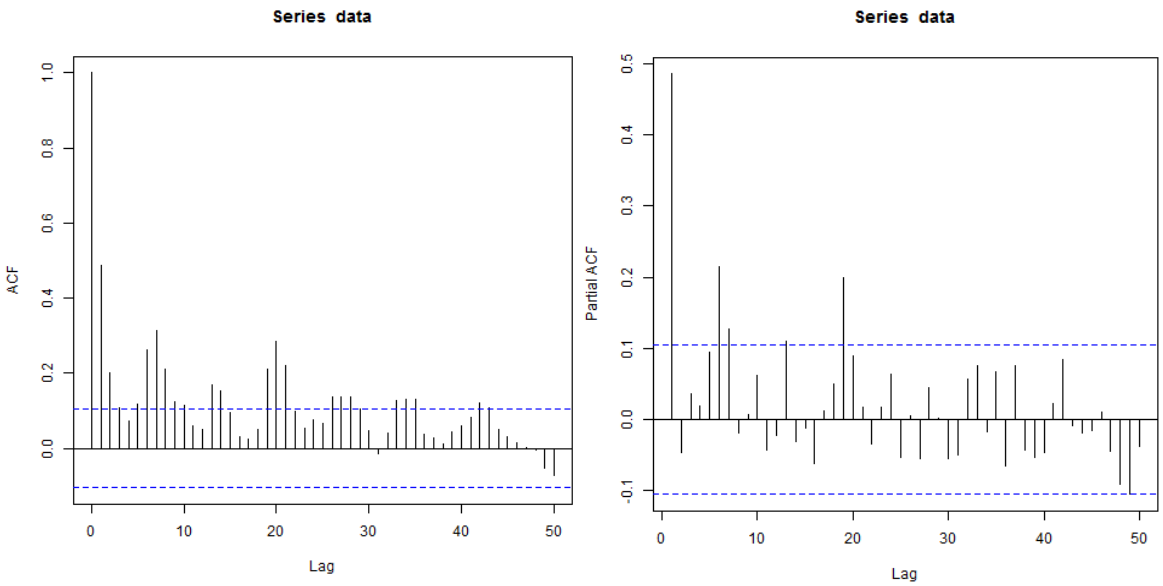

Figure 7 ACF and PACF results for CO_GT

There are some problems on linear model. One is that residuals cannot be considered as white noise. The Box-Pierce test shows that p-value is extremely close to 0, rejecting the null hypothesis that data are independently distributed. There are some sorts of autocorrelation which decreases the accuracy of the model. Meanwhile, the linear model is able to explain concentration for CO, but it is not capable of predicting. Thus I will turn to use some time series models to try to fix the autocorrelation problem and predict target variable.

4.3. AR model

Since co_gt is a stationary variable, AR model could be applied to it, as current concentration for CO might be affected by its previous values.

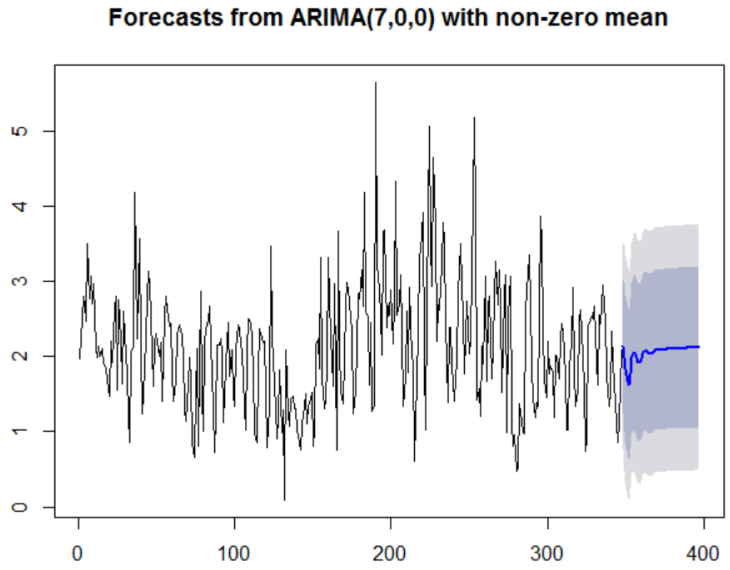

Figure 8 Predictions for CO concentration with AR(7) model

```
Call:
arima(x = data, order = c(7, 0, 0))

Coefficients:
         ar1      ar2     ar3      ar4      ar5     ar6     ar7  intercept
      0.4561  -0.0575  0.0264  -0.0161  -0.0059  0.1542  0.1299     2.1167
s.e.  0.0531   0.0579  0.0579   0.0580   0.0579  0.0579  0.0532     0.1178

sigma^2 estimated as 0.4898:  log likelihood = -368.92,  aic = 755.83

Training set error measures:
                      ME      RMSE       MAE       MPE     MAPE     MASE        ACF1
Training set -0.002866596 0.6998881 0.5259797 -17.12057 34.35972 0.840383 0.003596682
```

Table 5 Results for MA(7) model

Fig.7 shows the autocorrelation and partial autocorrelation for co_gt series. ACF converge to 0 gradually but we could hardly see a distinctive turning point where it approaches to 0 suddenly. However, for PACF, loosely speaking, it converges to 0 at the third order. Even though some are out of boundary, we think that it occurred by chance. As a result, AR model is more feasible than MA model. I picked AR(7) as the optimal model basing on AIC, and result is like Table 5.

The current CO concentration is negatively affected by ar2, ar4 and ar5, while positively affected by ar1, ar3, ar6 and ar7. Coefficients firstly decrease and then increase, and ar1 influence current CO concentration the most, showing that there exists a 7-day periodicity. It is not hard to explain the fact that while current daily average CO concentration is highly affected by the last value, it might also be similar to that of one week ago since there are some similar features on the same day within a week. Coefficients are all less than 1, indicating that the model is stationary.

The final model is:

$$\widehat{\text{co\_gt(t)}} = 0.4561co\_gt_{t-1} - 0.0575co\_gt_{t-2} + 0.0264co\_gt_{t-3} - 0.0161co\_gt_{t-4} - 0.0059co\_gt_{t-5} + 0.1542co\_gt_{t-6} + 0.1299co\_gt_{t-7} + 2.1167 + \varepsilon$$

Fig.8 shows the predicted CO concentrations in the next 50 days. It will be stable with some fluctuations. Results of Ljung-Box test indicates that residuals are white noise as the p-value is 0.2022, not rejecting the null hypothesis that residuals are independent and stochastic. From this perspective, performance of AR model is better than that of linear model in estimation. Also, the ACF result shows that there is little autocorrelation for residuals since it decreases sharply at the first lag order and then converge to 0.

4.4. ARMA model

As previously analysis stated, if loosely speaking, there is a turning point for PACF at the third order. However, if we think it more strictly, PACF shrinks but no distinctive turning points exist.

```
Call:
arima(x = data, order = c(7, 0, 12))

Coefficients:
         ar1     ar2      ar3     ar4     ar5      ar6     ar7     ma1      ma2     ma3     ma4      ma5     ma6     ma7
      0.4943  0.3450  -0.8665  0.8533  0.3887  -0.6517  0.3551  0.0045  -0.4418  0.7270  -0.528  -0.7599  0.6005  0.0373
s.e.  0.2201  0.1354   0.1635  0.1396  0.1539   0.1460  0.1851  0.2212   0.1315  0.1927   0.116   0.1337  0.2382  0.2040
          ma8     ma9    ma10     ma11     ma12  intercept
      -0.0594  0.0495  0.0570  -0.1163  -0.2130     2.1138
s.e.   0.1098  0.0869  0.0789   0.0678   0.0731     0.1477

sigma^2 estimated as 0.4385:  log likelihood = -353.02,  aic = 748.03
```

Table 6 Results for ARMA(6,12) model

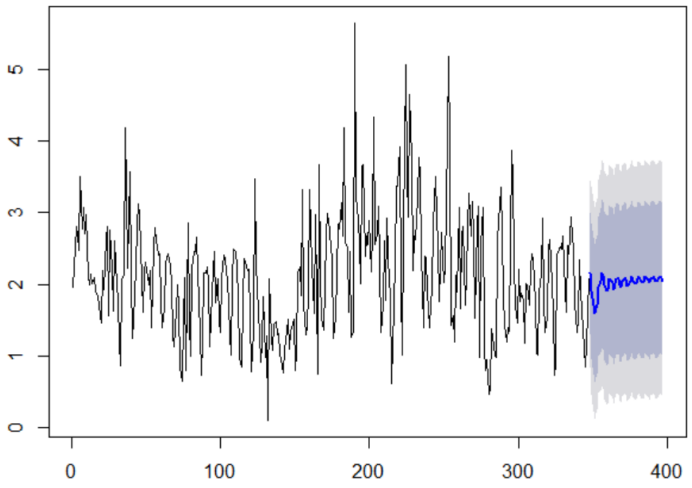

Figure 9 Predicted CO concentrations of ARMA model

In this case, when ACF and PACF converge to 0 gradually without a turning point, the ARMA model should be more effective.

Basing on AIC, ARMA(7,12) is chosen. Results are shown in Table 6. Coefficients for ar3, ar4, ar5 and ar6 significantly increase compared to AR model. All coefficients are less than 1, indicating that the model is stationary. When predicting future CO concentrations, there are more fluctuations, which reflect the effects of moving average operators. The ljung-box test result shows that mean of residuals is independent and stochastic, for the p-value is 0.9171, not rejecting the null hypothesis. ACF test for residuals also reflects that there is little evidences for residuals autocorrelation at 1-20 lags order. Residuals can be considered as completely white noise. These results are even better than AR model, since p-value is even largely greater than that of AR model and ACF shrinks more rapidly.

```
Estimation results for equation co_gt:
======================================
co_gt = co_gt.l1 + s2_nmhc.l1 + s4_no2.l1 + rh.l1 + s1_co.l1 + s3_nox.l1 + const

             Estimate Std. Error t value Pr(>|t|)
co_gt.l1    0.4234912  0.1117809   3.789 0.000179 ***
s2_nmhc.l1  0.0015941  0.0007712   2.067 0.039497 *
s4_no2.l1  -0.0005443  0.0002135  -2.549 0.011229 *
rh.l1       0.0095753  0.0035507   2.697 0.007352 **
s1_co.l1   -0.0004511  0.0005312  -0.849 0.396281
s3_nox.l1   0.0008480  0.0003265   2.597 0.009801 **
const      -0.1703558  0.7402239  -0.230 0.818121
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.7186 on 339 degrees of freedom
Multiple R-Squared: 0.2727,     Adjusted R-squared: 0.2598
F-statistic: 21.18 on 6 and 339 DF,  p-value: < 2.2e-16
```

Table 7 Results for VAR(1) model

4.5. VAR model

Though AR and ARMA models are both good to predict CO concentration, they did not involve exogenous variables. They just predict based on themselves. However, CO concentration is not affected just by itself. As section 4.2 analyzed, average CO concentration was affected by many factors significantly. So the involvement of exogenous variables might be important to the accuracy of a model.

The final variables for VAR model are s1_co, s2_nmhc, s3_nox, s4_no2 and relative humidity, which is exactly the same as linear model, after I tried different combination of factors and chose the one with the best significant level. Basing on BIC, I chose 1 as the lag order. Table 7 shows the results for VAR(1) model, regressing co_gt on exogenous factors.

The final VAR(1) model is:

$$\widehat{\text{co\_gt(t)}} = 0.4235co\_gt_{t-1} + 0.0016s2\_nmhc_{t-1} - 0.0005s4\_no2_{t-1} + 0.0009rh_{t-1} - 0.0005s1\_co_{t-1} + 0.0008s3\_nox_{t-1} - 0.1704 + \varepsilon$$

Co_gt.l1 is significant at 0.001% level. S3_nox.l1 and rh.l1 are significant at 0.01% level. The previous CO concentration still has the leading influence on current average CO concentration. Measured values for NMHC, NOx, NO2 and relative humidity also influence the average CO concentration a little bit to some degree. But the measured value for CO is not significant in this case. The F-statistic shows that these factors jointly influence target value significantly. Figure 10 shows the predicted CO concentration within the next 50 days. Compared to AR and ARMA model, the prediction is more stable with few fluctuations, but the trends are similar.

The p-value of the granger causality test was 0.00431, rejecting the null hypothesis that s2_nmhc, s4_no2, s1_co, s3_nox and relative humidity do not Granger-cause co_gt. The p-value

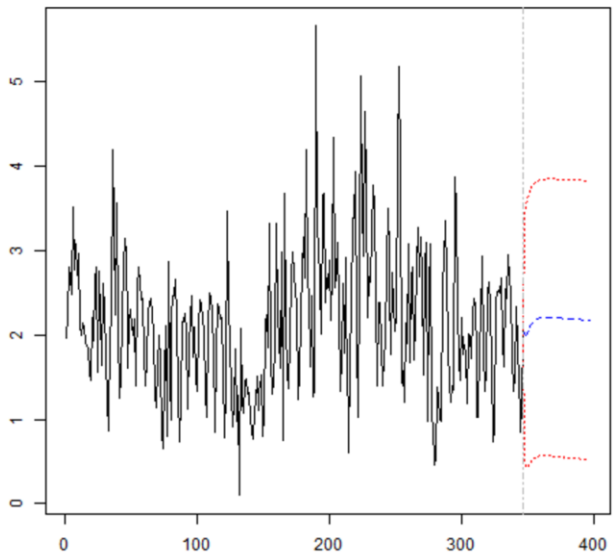

Figure 10 Predition for CO centration under VAR(1) model

of instant test was close to 0, rejecting the null hypothesis that there is no instantaneous causality between s2_nmhc, s4_no2, s1_co, s3_nox, relative humidity and co_gt. Both test consolidated the relationships between this variable combination and ground truth CO concentration.

4.6. Comparisons and problems

In previous analysis, I analyzed estimations of linear, AR, ARMA, VAR model, respectively. Generally speaking, they all have some advantages and disadvantages over other models.

The linear model interpret CO concentration well with the most optimal significant level, but it does not input the involvement of the time, basing on current states solely. Additionally, linear model can estimate but does not have the prediction function. AR, ARMA and VAR involve the participation of time, and are able to predict future trends.

AR and ARMA model only consider the involvement of CO concentration itself by including its previous values in the model. But linear regression and VAR model include exogenous factors. Since CO concentration itself is the leading factor, AR and ARMA model preform not worse than VAR model in predicting.

ARMA has the best capability to involve fluctuation in its prediction while AR and VAR model predict pretty stably. The ability to involve fluctuation makes the residuals more like white noise and has improved the accuracy indicated by Ljung box test.

There is a general problem that since it is a year-long data, we could not extract the seasonal features through any models, even though we could visually see that CO concentration might usually be higher and summer is lower.

5. Calibrate them or not?

Since I have analyzed characteristics and performance for several models, we have to come back to our calibration question. Obviously, the detected value for CO concentration is quite different from the trends of ground truth CO concentration. The multi sensor air quality device need to be calibrated based on the reference value and other endogenous estimators, but how should we apply them in the real world?

For estimating current ground truth CO concentration, though linear model is really good for it, there are some real restrictions when applying it. First of all, if we want to estimate the current ground truth CO concentration, we should firstly know the relative humidity and instant values of other four pollutants. But there is no relative humidity detector inside the multi sensor device. Considering implanting another humidity sensor would be a big risk for product design, for it might enlarge the volume of the device and increase the cost. We could consider gaining the value from official weather station, but as I mentioned, it is hysteresis, meaning that we could not get the value immediately so that the device lost its instantaneity. Fortunately, VAR(1) model should be a good choice. The device could calculate the current value of CO concentration based on previous data, so it is not necessary to consider the time delay. The result of VAR is quite reliable according to the Granger test.

For predicting ground truth CO concentration, ARMA model should be the best one. It includes more moving average parts, and it considered involvements of AR parts, too. There is no need to change the design of the device to get the reliable predictions.

With the help of statistical models, we could improve the accuracy of air quality multi sensor device for estimating and predicting CO concentrations. Bias has been fixed to some degree. It might be more practical for real world usage.

6. Conclusions and further works

This paper estimated and predicted CO concentration with various statistical models, including linear model, AR, ARMA and VAR model. From previous analysis, the current ground truth CO concentration is highly linear related to detected values for CO, NMHC, NO2, NOx by multi sensor device and relative humidity. Besides exogenous variables, CO concentration has some sorts of autocorrelation. The current ground truth concentration is affected by the last CO concentration the most, and then the one detected a week ago, indicating that there are some weekly patterns.

All models can be used to estimate CO concentration, but linear model cannot be used for prediction. Among the three time series models I used, ARMA model performs the best in predicting, since it allows more fluctuations in so that the result is more accurate. But it failed to interpret the relationships between ground truth CO concentration and other endogenous factors. The VAR(1) model includes involvement of endogenous factors, but it failed to include more previous values of CO concentration itself. Granger test shows that four measured values for CO, NMHC, NO2 and NOX, as well as relative humidity granger cause to ground truth CO concentration. Finally, VAR model was chosen to estimate the ground truth CO concentration and ARMA model was chosen to predict CO concentration, having calibrated the sensor CO concentration.

The work can be further improved. Firstly of all, we could consider to include both endogenous variables and more AR parts in the linear model. The current model includes either endogenous variables or several AR parts only. Secondly, cross terms could be considered. For example, NO2 and water can generate NOx, so it means that when auto emitted the same amount of pollutant, the higher-densed the water is(higher relative humidity), the lower the NO2 is and the higher NOx is. If we applied the result in linear model, CO concentration would increase, but the problem is CO was emitted the same. So if we multiply NO2 and relative humidity to generate a new cross term, it could better interpret CO concentration. Thirdly, cross validation can be introduced to test the performance of each model. We select some data as the training set and others as testing set. While fitting the model with training set, we can perform the model onto the testing set to compare the actual results and fitted values. Finally, in order to solve the seasonal change issue, we could perform structural break in estimating and predicting CO concentration. We could estimate it by seasons.

These models used successfully estimate and predict CO concentration, and can be applied to nose calibrations. It remained to be a future task to further improve the accuracy, decrease the

cost, optimize the test and consider the seasonal changes, for the purpose of making it more pragmatic for ordinary people.

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

Saverio De Vito, Marco Piga, Luca Martinotto, Girolamo Di Francia, CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, Sensors and Actuators B: Chemical, Volume 143, Issue 1, 4 December 2009, Pages 182-191, ISSN 0925-4005

S. De Vito, G. Fattoruso, M. Pardo, F. Tortorella and G. Di Francia, 'Semi-Supervised Learning Techniques in Artificial Olfaction: A Novel Approach to Classification Problems and Drift Counteraction,' in IEEE Sensors Journal, vol. 12, no. 11, pp. 3215-3224, Nov. 2012.