The O_3 NowCast: U.S. EPA's Method for Characterizing and Communicating Current Air Quality

Adam Reff,* David Mintz, and Liz Naess

U. S. Environmental Protection Agency, Research Triangle Park, NC 27711

E-mail: reff.adam@epa.gov

Background

Ambient ozone (O_3) is known to cause adverse human health effects, including aggravation of asthma and chronic obstructive pulmonary disease (COPD) (*1*). The U.S. Environmental Protection Agency (EPA) takes several measures to protect against such health effects, including setting national ambient air quality standards (NAAQS) and working with state, local and tribal agencies to implement national ambient air monitoring networks to determine NAAQS compliance (*1*). The forms of the NAAQS are typically multi-year statistics, but most monitors currently collect data on an hourly basis. Data is thus available which can potentially let people make real-time decisions that mitigate exposure. This is the basis of the AirNow website (https://www.airnow.gov/), which reports this hourly data to the public in real time.

EPA has been publicly reporting current air quality to the public since the mid 1970's to provide individuals with information they can use help mitigate human exposure. The modern reporting

tool is the Air Quality Index (AQI), which EPA uses to communicate daily air quality (*2*). The AQI uses color-coded categories and provides statements for each category that tell the public about air quality, which groups of people may be affected, and steps to take to reduce exposure to air pollution. It normalizes ambient concentrations of five pollutants (carbon monoxide, nitrogen dioxide, ozone, particulate matter, and sulfur dioxide) to an index value from 0 to 500. The daily AQI value is available the day after measurements are collected since a day's worth of complete data are needed for its calculation. Next-day forecasts of the AQI are also regularly made to allow the public to make informed decisions about their activity and potential exposure. However, as technology to measure short-term concentrations (1 hour or less) has improved in recent years, concern about the health impacts of shorter-term exposures has also grown. This is especially true during episodes of greatly varying air pollution concentrations, for example when forest fires are burning.

To address the need for real-time reporting of air pollution data to the public and because the AQI is based on longer-term averages, EPA has developed a "NowCast" to estimate the AQI in real time. For O_3, the NowCast is a prediction of what the AQI would be for an 8-hour average centered on the current hour. Centering on the current hour allows the NowCast to rise and fall with the current hourly concentration, while still representing an 8-hour average. It is a prediction of the AQI for the current hour ("Now"), because the conditions of the future hours are unknown. In this paper, we describe the methodology for computing the O_3 NowCast.

Methods

O_3 NowCast Calculation

The daily O_3 AQI is based on the highest 8-hour mean O_3 concentration over the course of a day (*3*). A real time O_3 AQI could thus be calculated if the 8-hour mean O_3 concentration representative of the current hour were known. The NowCast calculation we developed is therefore a prediction

of the 8-hour mean O_3 concentration from the 1-hour O_3 measurements.

This O_3 NowCast works by building a statistical relationship between 8-hour rolling mean O_3 concentrations and 1-hour concentration measurements. This relationship is then applied to the current 1-hour measurement to predict the 8-hour mean. The AQI value that corresponds with this 8-hour mean can be computed to provide a message to the public in real time about air quality and exposure reduction measures. The details of this statisical relationship are provided below.

This method of performing the NowCast takes inspiration from the idea of autoregression (AR) in time series analysis (*4*) in the sense that a regression model is developed between current and previous values of a data stream generated at regular time intervals. The regression method called partial least squares (PLS) is used to relate 8- and 1-hour concentrations of O_3. The `pls` package in R statistical software (*5*) is used to perform these calculations. PLS is a well established method for instrument calibration in analytical chemistry, where unknown concentrations of samples are predicted from instrument signals (*6*). The general equation for PLS is in the form of a regression equation:

$$y = \sum_{i=1}^{n} \beta_i \cdot x_i \tag{1}$$

However PLS differs from normal linear regression by employing principles of factor analysis in the algorithm to optimize results for prediction rather than a descriptive characterization.

Using this method, y is the 8-hour mean O_3 concentration centered on the current hour. The x_i are the 1-hour concentrations for both the current and $n-1$ previous hours' 1-hour O_3 concentration measurement. For this work, a pair of nested rolling windows is employed to apply this model to the data. The outer rolling window is the 2-week period of hours (336) that ends at the current hour. The inner window is a 4 day window of hours (n = 96) which assembles the 8-hour and 1-hour O_3 concentrations into a block of data to which PLS will be applied. This block of data

has a format as follows:

$$\begin{array}{ccccccccc}
\beta_1 \cdot x^1 & + & \beta_2 \cdot x^2 & + & ... & + & \beta_{96} \cdot x^{96} & = & y^{96} \\
\beta_1 \cdot x^2 & + & \beta_2 \cdot x^3 & + & ... & + & \beta_{96} \cdot x^{97} & = & y^{97} \\
... & & & & & & & & \\
\beta_1 \cdot x^{239} & + & \beta_2 \cdot x^{240} & + & ... & + & \beta_{96} \cdot x^{335} & = & y^{335}
\end{array} \tag{2}$$

Note that the subscripts are referring to the index of hour in the inner window (4 days wide) while the superscripts are the index of the hour in the outer window (2 weeks wide). The PLS model is applied to the data arranged in this way in order to find the β_i values that relate each possible 4 day sequence of hourly O_3 values (x_1 through x_{96}) to the 8-hour O_3 value on the right side of the inner window (y). This gives a set of β_i representative of the 2 week window that ends at the current hour. These β_i values are then applied to x^{240} through x^{336} to calculate a prediction for the 8-hour O_3 mean at the current hour (y^{336}). This process is repeated in real time each time a new hourly O_3 measurement is obtained. The sizes of the 2 rolling windows were obtained through trial and error to find values that balanced accuracy and computation time.

Data Handling

Limitations in the O_3 monitor and measurement logistics sometimes lead to unavailable data. Procedures for minimizing the effect of this missing data on the NowCast were thus incorporated into the algorithm. Hourly concentrations can be missing entirely (labeled as "NA", Not Available), or can be ≤ 0. The following completeness criteria are thus implemented in order to calculate a valid NowCast:

1. Each 8 hour window must be 75% complete (6 values) to calculate an 8-hour mean, otherwise the 8-hour mean is NA.

2. PLS requires no missing values in either the X or y sides of Equation 2. The following steps are thus taken to ensure there are no missing values:

    - A new imputed version of the 1-hour O_3 data is created to make the X block of Equation 2 complete. Note that this imputed version of the data is not used for calculation the rolling 8-hour means used in the y side of Equation 2. Imputation is done through the moving average method present in the `na.ma()` function of the `imputeTS` R package (*7*). This imputation is only performed if the 2-week stream of data contains no sequences of missing data longer than 7 hours.
    - Rows containing NAs in the y variable (as a result of not meeting Condition 1 above) are removed from the PLS regression entirely. No more than 25% of rows can be removed, else a surrogate (see below) NowCast is computed.

3. The current and previous 2 hours must not be NA.

4. The resulting NowCast calculation must be > 0.

If all of the above conditions are met, the NowCast dervied from the PLS method is used and reported to the public. To handle cases where the above conditons are not met, the following steps are taken:

1. If all data are missing for the current 2 week period, the NowCast is set to NA.

2. If all data are 0 for the current 2 week period, the NowCast is set to 0.

3. If the resulting NowCast is < 0, the NowCast is set to 0.

4. If the current hour is missing an O_3 measurement, the previous hour's NowCast is used.

5. If the current and previous hours are missing O_3 measurements, the NowCast from 2 hours ago is used.

6. If the latest 3 hours are missing concentrations, the current NowCast is set to NA.

7. If none of the above issues are present, but there is 1) insufficient data for 75% completeness of 8-hour means, or 2) a sequence of missing 1-hour O_3 concentrations longer than 8 hours, then an alternative calculation is used. The NowCast is calculated from the following equation:

$$NowCast_i = 0.85 * [O_3]_i + 4.5 \tag{3}$$

The NowCast from hour i is then a simple function of the 1-hour O_3 concentration from the same hour i. This equation was derived from a linear regression between concurrent 8-hour mean and 1-hour O_3 concentrations using historical data from about 40 monitoring sites from major continental U.S. cities.

Code, Sample Data, and Outputs

Data providers are encouraged to use the computed NowCast from AirNow which is available using the AirNow API (`https://docs.airnowapi.org/`). However, we have provided some resources for those who want to compute the O_3 NowCast themselves on GitHub at the following link: `https://github.com/USEPA/O3-NowCast/tree/master`. The following items can be found there:

1. R Code to perform the NowCast calculation

2. An example input data

3. The output file generated using the example input

4. Example graphics of the inputs & outputs

Literature Cited

(1) EPA, *Review of the National Ambient Air Quality Standards for Ozone: Policy Assessment of Scientific and Technical Information*; Technical Report EPA-452/R-07-007, 2007.

(2) EPA, *Technical Assistance Document for the Reporting of Daily Air Quality - the Air Quality Index (AQI)*; Technical Report EPA-454/B-18-007, 2018.

(3) Mintz, D. *Guideline for Reporting of Daily Air Quality - Air Quality Index (AQI)*; Technical Report EPA-454 B-06-001, 2006.

(4) Greene, W. H. *Econometric Analysis*; Prentice Hall, 2003.

(5) Mevik, B.-H.; Wehrens, R.; Liland, K. H. *pls: Partial Least Squares and Principal Component regression*; 2011, R package version 2.3-0.

(6) Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley and Sons: New York, 1989.

(7) Moritz, S.; Bartz-Beielstein, T. imputeTS: Time Series Missing Value Imputation in R. *The R Journal* **2017**, *9*, 207–218.