Atmospheric Environment 103 (2015) 53–65

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Real time air quality forecasting using integrated parametric and non-parametric regression techniques

Aoife Donnelly*, Bruce Misstear, Brian Broderick

Department of Civil, Structural and Environmental Engineering, Museum Building, Trinity College Dublin, College Green, Dublin 2, Ireland

HIGHLIGHTS

- A statistical model to provide real time hourly forecasts of NO_2 is presented.
- Non-parametric kernel regression is applied in parallel with multiple linear regression.
- The model has low computational resources and requires simple input data.
- IA values of between 0.74 and 0.94 were obtained.

ARTICLE INFO

Article history:
Received 25 June 2014
Received in revised form
24 October 2014
Accepted 4 December 2014
Available online 5 December 2014

Keywords:
Nitrogen dioxide
Nonparametric kernel regression
Air quality forecasting
Statistical modelling

ABSTRACT

This paper presents a model for producing real time air quality forecasts with both high accuracy and high computational efficiency. Temporal variations in nitrogen dioxide (NO_2) levels and historical correlations between meteorology and NO_2 levels are used to estimate air quality 48 h in advance. Non-parametric kernel regression is used to produce linearized factors describing variations in concentrations with wind speed and direction and, furthermore, to produce seasonal and diurnal factors. The basis for the model is a multiple linear regression which uses these factors together with meteorological parameters and persistence as predictors. The model was calibrated at three urban sites and one rural site and the final fitted model achieved R values of between 0.62 and 0.79 for hourly forecasts and between 0.67 and 0.84 for daily maximum forecasts. Model validation using four model evaluation parameters, an index of agreement (IA), the correlation coefficient (R), the fraction of values within a factor of 2 (FAC2) and the fractional bias (FB), yielded good results. The IA for 24 hr forecasts of hourly NO_2 was between 0.77 and 0.90 at urban sites and 0.74 at the rural site, while for daily maximum forecasts it was between 0.89 and 0.94 for urban sites and 0.78 for the rural site. R values of up to 0.79 and 0.81 and FAC2 values of 0.84 and 0.96 were observed for hourly and daily maximum predictions, respectively. The model requires only simple input data and very low computational resources. It found to be an accurate and efficient means of producing real time air quality forecasts.

1. Introduction

Air quality forecasts are required in the European Air Quality Directive in instances where concentrations exceed or are expected to exceed alert and information thresholds (EEA, 2011). Such models need to be capable of being run routinely with minimum resource requirements. Routine air quality forecasts are of high importance from a public health, air quality management and scientific perspective. Densely populated areas and urban locations benefit significantly from air quality forecasting as population warnings and emergency control measures can be implemented in advance of pollution episodes. These forecasts should necessarily be available 24–48 h in advance of the episode. Nitrogen dioxide (NO_2) is one of the main pollutants of concern (Environmental Protection Agency, 2012) and varies temporally and spatially with anthropogenic emissions and meteorological conditions (Jensen, 1998). Emissions due to transport or fossil fuel combustion depend on human activity but their effects on concentrations at a particular receptor are also influenced by meteorology and the nature of the receptor. The current study is concerned with producing 24 and 48 h forecasts of hourly and daily maximum NO_2 at rural and urban sites.

* Corresponding author.
E-mail address: aadonnel@tcd.ie (A. Donnelly).

http://dx.doi.org/10.1016/j.atmosenv.2014.12.011

Statistical modelling has been found by many countries to offer a viable and attractive alternative to large scale deterministic models when developing operational air quality modelling capabilities (e.g. Lissens et al., 2000; Chaloulakou et al., 2003; Cobourn, 2007). Like deterministic models, statistical models tend to be comprised of different smaller models. A major advantage of statistical models is that they can be developed from first principles specific to the area of interest, removing reliance on third party model suppliers. Zhang et al. (2012) in their recent review of real time air quality forecasting systems note that while statistical approaches generally require a large quantity of historical measured data under a variety of conditions, they often have higher accuracy when compared to deterministic models.

This study presents an elegant model which requires minimal computing facilities for the prediction NO_2 concentrations out to 48 h. The model is created by combining a time series model (parametric and non-parametric), a nonparametric kernel regression model and a multiple linear regression model. Emission sources are represented by temporal concentration profiles produced by the nonparametric kernel regression model, removing any requirement for an emissions inventory.

2. Methodology

2.1. Calculation

The basis for the air quality prediction is a multiple linear regression (MLR) which uses as inputs:

- Linear factors generated from a non-parametric kernel regression model
- Forecast meteorological parameters.

The method builds on previous research by the authors which applied a two dimensional non-parametric kernel regression technique to quantify the effects of wind direction and speed on background NO_2 concentrations (Donnelly et al., 2011, 2012). In parametric regression, sample data are used to estimate the values of the regression coefficients. Such regression is linear if the response variable is assumed to be a linear function of the regression coefficients. Previous work by the authors found that the variation in NO_2 concentration levels with wind speed and direction was nonlinear but was well described using non-parametric kernel regression methods. Nonparametric regression relaxes the functional form assumed in parametric regression, the object being to estimate the regression function directly, rather than to estimate parameters (Donnelly et al., 2011). A further distinguishing feature of nonparametric regression is "the nonexistence of an inclination to reduce the number of parameters in the equation" (Takezawa, 2005).

Donnelly et al. (2011) tested a powerful tool for the quantification of the effects of wind direction and speed on background NO_2 concentrations, particularly in cases where monitoring data are limited. In contrast to frequently used methods such as data binning, nonparametric regression allows concentrations values in missing data pairs to be estimated and distinction between spurious and true peaks in concentrations to be made. Accurate identification of the actual variation at each location and causative factors could be made, thus supporting the improved definition of concentrations for use in air quality modelling studies. The output from the regression is a set of linearized factors for each wind speed/wind direction pair which can then be used as inputs to the MLR.

The model development, calibration and validation is described in the following section and it is helpful to read this in conjunction with Fig. 1. The general form of the model is:

$$C = b_0 + \sum_{i=1}^{n} b_i x_i + \sum_{i=1}^{m} d_i y_i + \varepsilon$$

where C is the response variable (NO_2 concentration), b_0 is the regression constant, the x_i are the meteorological predictor variables with coefficients b_i, and the y_i are the predictor variables output from the non-parametric and time series models with coefficients d_i. ε is the stochastic error associated with the regression. A least squares technique was used to determine the coefficients for each of the following predictor variables:

y_i - Factors developed for each site
- The wind speed, wind direction factor as output from the nonparametric regression ($WSWD_f$)
- Non-parametric seasonal factor (S_f)
- Non-parametric diurnal factor (D_f)
- Time series forecast factor (TS_f)

x_i - Variables measured at each site
- Hourly temperature, Sunshine, Relative humidity, Atmospheric pressure, Stability class
- Hourly NO_2 concentration at 24 or 48 h lags (NO_2h-24, NO_2h-48)
- Daily average NO_2 concentration at 24 or 48 h lags (NO_2d-24, NO_2d-48)
- Daily maximum NO_2 concentration at 24 or 48 h lags (NO_2max-24, NO_2max-48)
- Daily average O_3 concentration at 24 or 48 h lags (O_3d-24, O_3d-48)
- Daily minimum O_3 concentration at 24 or 48 h lags (O_3min-24, O_3min-48)

$WSWD_f$ was developed using non-parametric kernel regression (as described in Yu et al., 2004; Donnelly et al., 2011) and is calculated as follows:

$$WSWD_f = \frac{\tilde{C}(\vartheta, u, h, \gamma)}{\overline{C}}$$

Where $\overline{C}$ is the average concentration for the entire time series and $\tilde{C}(\vartheta, u, h, \gamma)$ is the average concentration of a pollutant for a given wind direction/speed pair (ϑ, u) calculated as a weighted average of the data in a window (of width defined by smoothing parameters h and γ using weighted kernel function $K(\vartheta, u, h, \gamma) = K_1(\vartheta, h) K_2(u, \gamma)$ around (ϑ, u) and defined as follows:

$$\tilde{C}(\vartheta, u, h, \gamma) = \frac{\sum_{i=1}^{N} K_1\left(\frac{(\vartheta - W_i)}{h}\right) K_2\left(\frac{(u - U_i)}{\gamma}\right) C_i}{\sum_{i=1}^{N} K_1\left(\frac{(\vartheta - W_i)}{h}\right) K_2\left(\frac{(u - U_i)}{\gamma}\right)}$$

where C_i, W_i and U_i are the observed concentration of a particular pollutant, resultant wind direction and speed for the ith observation in a time period starting at time t_i. For circular data such as wind direction the Gaussian kernel (K) is the preferred method used to weight the observations (Henry et al. 2002) and is defined as follows:

$$K_1(x) = (2\pi)^{-1/2} \exp\left(-0.5x^2\right) \quad -\infty < x < \infty$$

The Epannechnikov kernel is used for wind speed as it is the simplest bounded kernel (Yu et al. 2004):