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

Aoife Donnelly, Bruce Misstear, Brian Broderick

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 (NO₂) levels and historical correlations between meteorology and NO₂ 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 NO₂ 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.

Keywords

Nitrogen dioxide; Nonparametric kernel regression; Air quality forecasting; Statistical modeling

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 (NO₂) 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 NO₂ at rural and urban sites.

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

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 NO₂ 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 NO₂ 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 NO₂ concentration levels with wind speed and direction was non-linear 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 NO₂ concentrations, particularly in cases where monitoring data are limited. In contrast to frequently used methods such as data binning, non-parametric 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:

$$C = b_0 + \sum_{i=1}^n b_i x_i + \sum_{j=1}^m d_j y_j + \varepsilon$$

where C is the response variable (NO₂ concentration), b_0 is the regression constant, the x_i are the meteorological predictor variables with coefficients b_i , and the y_j are the predictor variables output from the non-parametric and time series models with coefficients d_j . 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_j - Factors developed for each site

- The wind speed, wind direction factor as output from the nonparametric regression (WSWD_{*j*})
- Non-parametric seasonal factor (S_j)
- Non-parametric diurnal factor (D_j)
- Time series forecast factor (TS_{*j*})

x_i - Variables measured at each site

- Hourly temperature, Sunshine, Relative humidity, Atmospheric pressure, Stability class
- Hourly NO₂ concentration at 24 or 48 h lags (NO₂h-24, NO₂h-48)
- Daily average NO₂ concentration at 24 or 48 h lags (NO₂d-24, NO₂d-48)
- Daily maximum NO₂ concentration at 24 or 48 h lags (NO₂max-24, NO₂max-48)
- Daily average O₃ concentration at 24 or 48 h lags (O₃d-24, O₃d-48)
- Daily minimum O₃ concentration at 24 or 48 h lags (O₃min-24, O₃min-48)

WSWD_{*j*} was developed using non-parametric kernel regression (as described in Yu et al., 2004; Donnelly et al., 2011) and is calculated as follows:

$$\text{WSWD}_j = \frac{\bar{C}(\vartheta, u, h, \gamma)}{\bar{C}}$$

Where $\bar{C}$ is the average concentration for the entire time series and $\bar{C}(\vartheta, u, h, \gamma)$ 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:

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

where C_i , W_i and U_i are the observed concentration of a particular pollutant, resultant wind direction and speed for the i th 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:

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

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

Fig. 1. Methodology flowchart.

$$K_2(x) = 0.75(1 - x^2) \quad \text{for } -1 \leq x \leq 1 \quad \text{and} \quad K_2(x) = 0 \quad \text{otherwise}$$

As discussed in Silverman (1986) a bandwidth of $0.9\sigma n^{-1/5}$ was employed, where σ is the standard deviation of the predictor variable data (wind speed or direction) and n is the number of data points.

Time series analysis was performed on the raw data from each site to produce TS_f . A seasonal multiplicative model was constructed as follows. Firstly a trend line was fitted to the data using least squares regression (the response is hourly concentration (NO_{2h}) and the predictor will be the observation number (t)). The resulting equation is:

$$NO_{2h} = a + b \times t$$

where a is the concentration at $t = 0$ and b is the trend in the data. The raw data are de-trended by dividing out the trend component as predicted from the above equation. The de-trended data are then smoothed using a centred moving average with length equal to the

seasonal cycle (8760 h). The median raw seasonal concentration is found within each seasonal period. This produces seasonal indices which are used to seasonally adjust each year's data. Using the trend and the seasonal components forecasts are made. In this case a full year of forecasts is made (8760) for the validation year.

S_f is derived using the non-parametric technique described above. This is a circular regression, i.e. the concentration in month 12 influences the concentration in month 1. The normalised seasonal factors $\left(\frac{\hat{C}(\alpha, h)}{\bar{C}}\right)$ are obtained after de-trending. Final S_f are obtained by scaling up these normalised factors and applying the relevant long term trend.

$$S_f = a + \frac{\bar{C}(\alpha, h)}{\bar{C}} \times b \times t$$

where a , b and t are as before, $\bar{C}$ is the average concentration for the input data used in model development and $\hat{C}(\alpha, h)$ is the average concentrations of a pollutant for a given day of the year (α) calculated as a weighted average of the data in a window (of width defined by smoothing parameter h using weighted Gaussian kernel function $K_1(\alpha, h)$ around (α) and defined as follows:

$$\tilde{C}(\vartheta, h_i) = \frac{\sum_{i=1}^N K_1\left(\frac{(\alpha - S_i)}{h}\right) C_i}{\sum_{i=1}^N K_1\left(\frac{(\alpha - S_i)}{h}\right)}$$

where C_i are de-trended concentrations, S_i is the day of the year for the i th observation in a time period starting at time t_i . The bandwidth is calculated based on the number of days in a year.

In developing the D_f the data are first subdivided into four categories distinguishing between winter and summer, and between weekdays and weekends. The resulting factors are developed in exactly the same way as S_f but in this instance hours are used in replacement of days (i.e. S_i is replaced by H_i where H_i is the hour of the day).

2.2. Model performance evaluation

Model performance evaluation is an area of much discussion and standard evaluation procedures still do not exist. Various statistical performance methods have differing strengths and weaknesses and therefore in this study, as discussed in Willmott (1982) a number of methods have been applied to compare modelled and monitored data: R (Correlation coefficient), IA (Index of Agreement), the fraction of predictions within a factor of two of observations (FAC2) and the fractional bias (FB). Since the bias in a multiple linear regression model will always be zero this parameter has only been assessed for the validation data.

$$R = \frac{(\overline{C_o} - \overline{C_p})(\overline{C_p} - \overline{C_p})}{\sigma_{C_p} \sigma_{C_o}}$$

$$IA = \frac{(\overline{C_p} - \overline{C_o})^2}{(|\overline{C_p} - \overline{C_o}| + |\overline{C_o} - \overline{C_o}|)^2}$$

$$FAC2 = \text{fraction of data that satisfy } 0.5 \leq \frac{C_p}{C_o} \leq 2$$

$$FB = \frac{(\overline{C_o} - \overline{C_p})}{0.5(\overline{C_o} + \overline{C_p})}$$

where, C_p are the model predictions, C_o are the observations, $\overline{C}$ is the average over the data set, σ_c and σ_p are the standard deviations of the observed and predicted data sets, respectively. The reader is referred to the following studies where some of these techniques were applied to assess air quality model performance (Elbir, 2003; Karppinen et al., 2000; Gokhale and Raokhande (2008; Voukantsis et al. 2011).

This study presents two distinct applications of model evaluation. The first is *model calibration* whereby the model is developed using historical data. This is an iterative process which uses the VIF together with graphical techniques and normality assessment to reach the final model. This final model is then assessed using model performance methods. The second application is *model validation* whereby the calibrated model is used to produce a set of independent concentrations which are then compared with measured values.

3. Data

EU legislation requires that member states divide their territory into zones for the assessment and management of air quality

(O'Dwyer, 2012). In 2012 in Ireland these are Dublin City (Zone A), Cork City (Zone B), large towns (populations over 15,000) (Zone C) and rural areas (Zone D). The model described in this paper has been developed for four sites as shown in Fig. 2, three in the Dublin zone and one in the rural zone. The data used in the research have been collected by the Irish Environmental Protection Agency (EPA) who are responsible for maintaining the air quality monitoring network in Ireland.

Winetavern Street is located in Dublin city centre and is classified as an urban centre site. Rathmines is classified as an urban background site and located in a residential and commercial area to the south of the city centre. Dun Laoghaire located further south of Dublin city centre in a residential region and is a suburban background site. The main emission sources at each of these sites are commercial/residential combustion and traffic emissions. The final site, Kilkitt is a rural background site. This site is remotely located in an agricultural region with no major local emission sources. NO_2 is monitored and recorded at hourly resolution at each site using chemiluminescence samplers (as recommended for demonstration of compliance with EU limit values (CEU, 2008)). NO_2 is indirectly measured by reducing to NO by means of a molybdenum converter. The instrument has a range of 0–20,000 ppb and a lower detection limit of 0.4 ppb with a precision of 0.5% of the reading. Average concentrations at all sites are well within an appropriate range for the instrument used.

In this paper the model is developed and calibrated using monitoring data from 2007 to 2011 (inclusive) hereby referred to as *calibration data*. In order to test the model, data from the most recent year were not used in model development and are hereby referred to as *validation data*. The calibrated model was used to predict concentrations at each site for a week from each season at all of the sites (chosen using stratified random sampling) as follows:

- 14/03/2012–21/03/2012
- 15/07/2012–22/07/2012
- 08/10/2012–15/10/2012
- 19/12/2012–26/12/2012.

Fig. 3 shows the long run mean NO_2 concentrations recorded at each site and indicates the data used for calibration and that used for validation.

4. Results

4.1. Wind speed/direction factor

Figs. 4–7 show $WSWD_f$ as produced by the non-parametric regression model for each season and site. At the three urban sites the highest $WSWD_f$ is predicted for low wind speeds. At both Winetavern Street and Rathmines there is a peak for all winds in a north-easterly/easterly direction. At both Winetavern Street and Dun Laoghaire a further peak is observed for north-westerly winds. All of these peaks correspond with the local road network in each area indicating that traffic emissions are the main contributors to NO_2 concentrations. $WSWD_f$ shows more absolute variation at the rural site than at the urban site. Highest concentrations are observed for moderate wind speeds in both seasons which is thought to be due to the absence of any major sources in the locality and the requirement for moderate winds to transport emissions from a road located over 12 km from the site.

4.2. Seasonal variations

Fig. 8 shows S_f for each site as output from the non-parametric seasonal model. The degree of seasonal variation appears to be

Fig. 2. Monitoring site locations.

Fig. 3. Yearly evolution of NO_2 at each site.

inversely related to the anthropogenic influence at the site, i.e. the most urbanised site (Winetavern Street) displays the lowest relative seasonal variation, while the most rural site (Kilkitt) displays the highest relative seasonal variation. Winetavern Street is located in the urban centre of Dublin where traffic volumes remain high all year round. Anthropogenic emissions appear to be the controlling factor on concentrations at this site and while seasonal meteorological factors serve to decrease concentrations slightly during summer months through better dispersion conditions the effect is minor compared to that at more rural sites. At Kilkitt, there is little anthropogenic influence in the locality and concentrations in January are approximately double those in July. The seasonal effect is slightly less pronounced at Rathmines and Dun Laoghaire due to an increased year-round anthropogenic influence.

4.3. Diurnal variations

Fig. 9 shows D_f at each site as output from the non-parametric diurnal time series model. NO_2 concentrations at Rathmines and

Fig. 4. WSWD_f at Winetavern Street winter/summer.

Fig. 5. WSWD_f at Rathmines winter/summer.

Fig. 6. WSWD_f at Dun Laoghaire winter/summer.

Fig. 7. WSWD_f at Kilkitt winter/summer.

Fig. 8. S_f as modelled at each site.

Dun Laoghaire follow a similar diurnal pattern during winter months. Concentrations are lowest at 5am on weekdays and at 6am on weekends. During the week there is a sharp increase due to rush hour traffic emissions. Peak concentrations at Dun Laoghaire tend to lag those of the urban centre sites by approximately 1 h due to its suburban location, lower peak traffic volumes and subsequent transport of emissions from city centre locations under certain wind conditions.

The morning peak at Winetavern Street is more prolonged than at the other urban sites. There is also less of an afternoon trough due to continuously high traffic volumes in the area during the day. The superposition of the rush hour peak on already elevated concentrations arising due to surrounding and outlying roads thus creates higher overall concentrations but less extreme variation.

There is less variation at Kilkitt than at the urban sites due to the lack of anthropogenic sources in the area. Weekends in both seasons exhibit a gradual increase towards a peak which occurs later in summer months than in winter months. Weekdays show a morning increase followed by a levelling off in the afternoon (due to improved mixing conditions) with a further evening peak (occurring earlier in winter months than summer months due to poorer mixing conditions and an earlier sunset). Other studies have

illustrated similar diurnal patterns in NO_2 levels at both urban and rural sites (e.g. Martin and Barber, 1981; Mazzeo et al., 2005) while studies at lower latitudes showed less seasonal change in diurnal variation due to a more consistent day length (e.g. Mavroidis and Iliu, 2012).

4.4. Multiple linear regression

A stepwise process was used to isolate the significant explanatory variables in the MLR. At each site the predictor variables were examined for co-linearity and for inclusion in the final model. Multicollinearity increases the standard errors of the coefficients which means that coefficients for some independent variables may be erroneously found significant. The variance inflation factor (VIF) was used to assess how much the variance of an estimated regression coefficient increases if predictors are correlated; it is equal to 1 if no factors are correlated. Variables with high VIF were removed from the model in a stepwise manner ensuring that each variable removed is redundant in the explanation of NO_2 concentration. Definition of the final model also relied on knowledge of the directions of influence for particular parameters and graphical techniques.

In each case WSWD_f , D_f , S_f were significant variables. In general it was found that $\text{TS}_{\text{factor}}$ should not be included in the model when S_f and D_f are used and the use of the latter was found to be preferable. WSWD_f was a key explanatory variable at all sites and clear peaks and troughs for certain wind speed and direction pairs could be associated with emissions in the area. The coefficient is of similar magnitude at both Rathmines and Winetavern Street. It is lower at Kilkitt as WSWD_f itself had a higher relative variation here due to its location distant from any anthropogenic sources. The D_f coefficient is of similar magnitude at each site as WSWD_f . When anthropogenic sources dominate, S_f does not carry a high regression coefficient, emphasised by the most urban site (Winetavern Street) having the lowest coefficient. S_f has a high coefficient at Kilkitt, most likely due to the natural influences on concentrations at the

Fig. 9. D_f weekdays/weekends (left/right) and winter/summer (top/bottom).

Table 1
Model evaluation parameters for hourly and daily maximum predictions (calibration data).

		Winetavern		Rathmines		Dun Laoghaire		Kilkitt	
		24 h	48 h	24 h	48 h	24 h	48 h	24 h	48 h
Hourly	R (adj)	0.76	0.73	0.79	0.78	0.69	0.68	0.62	0.58
	IA	0.84	0.83	0.88	0.87	0.78	0.78	0.73	0.69
	FAC2	0.81	0.79	0.81	0.79	0.61	0.60	0.55	0.52
Daily maximum	R (adj)	0.76	0.71	0.84	0.83	0.77	0.76	0.67	0.61
	IA	0.84	0.81	0.90	0.90	0.86	0.85	0.77	0.72
	FAC2	0.97	0.96	0.96	0.96	0.90	0.90	0.71	0.70

Fig. 10. Modelled vs measured hourly NO₂ at: Top-Winetavern Street/Rathmines, Bottom-Dun Laoghaire/Kilkitt.

rural site and the lower degree of explanation provided by the other model parameters.

NO₂h-24 and O₃d-24 were significant at all sites and since anthropogenic activity is less likely to influence concentrations at the rural site NO₂h-24 was found to carry the highest relative coefficient here. The previous day's ozone concentrations were negatively correlated with hourly NO₂ concentrations at the rural site as, in the absence of local anthropogenic (and VOC) emissions, meteorological conditions tend to control the relationship between NO₂ and ozone. The conditions that lead to low ozone levels tend to result in higher NO₂ levels and this tends to persist from one day to the next. Other studies have frequently shown a negative correlation between NO₂ and Ozone (e.g. (Abdul-Wahab et al. 2005)). At urban sites weak (positive) correlations were observed between NO₂ and ozone. The presence of VOCs can lead to the formation of additional ozone and as a result concurrent high concentrations of NO₂ and ozone can be found at sites where VOCs are also emitted. In contrast the rural site has very low VOC levels.

Stability class was significant at all sites. Relative humidity was insignificant at both Rathmines and Winetavern Street and had low importance at Dun Laoghaire and Kilkitt. Air pressure has a similar low importance at Rathmines and Winetavern Street and was insignificant at Dun Laoghaire but was significant at Kilkitt. This

latter positive correlation is most likely due to warming sinking air associated with high pressure regions that can create upper layer inversions, light surface winds and low cloud cover. Temperature has a weak negative correlation at Winetavern Street and Kilkitt but was co-linear with S_f and not a useful addition to the model. Sunshine is an important parameter at Dun Laoghaire (attributed to the coastal location of this site) but was insignificant at Winetavern and Kilkitt and of low importance at Rathmines.

4.5. Model calibration

In this section the model coefficients described above are used to produce modelled values which are compared to measured data using model performance methods. Table 1 presents the model fitting parameters of the final fitted models for hourly and daily maximum concentrations. The highest explanation of total variation is achieved at Rathmines. This site also has the lowest relative error. The lowest amount of variation is explained at the rural site (Kilkitt).

IA is high at all sites, reaching a value of 0.88 at Rathmines for 24 h predictions and 0.87 for 48 h predictions. Highest R values are observed at the two urban sites (0.76–0.79 for 24 h forecasts). At the suburban and rural sites this is slightly lower (0.69 and 0.62,

Fig. 11. Modelled vs measured daily maximum NO₂ at: Top-Winetavern Street/Rathmines, Bottom-Dun Laoghaire/Kilkitt.

respectively). FAC2 is highest at the two urban sites, varying between 0.79 and 0.81. It is slightly lower at the suburban site (between 0.60 and 0.61). At the rural site it is 0.55 for 24 h predictions and 0.52 for 48 h predictions.

Fig. 10 shows hourly modelled concentrations versus measured concentrations at the four sites. At the three urban sites low to moderate concentrations are well predicted. In general there is some under prediction of high pollution events. Predictions are poorest at the rural site due to the very low concentrations and, furthermore, the lower influence of anthropogenic factors means that there is less consistency in the variations.

To capture the peak concentrations observed on a daily basis, the daily maximum NO₂ concentration was used as the predictor variable (Fig. 11). *R* values are higher at urban and suburban sites between 0.76 and 0.84 for 24 h predictions and between 0.71 and 0.83 for 48 h predictions. The index of agreement is high across all sites (up to 0.90 at urban sites and 0.77 at rural sites). FAC2 is very high at urban sites (up to 0.97) and significantly higher for this model at the rural site than the hourly model (0.71). These results indicate that the daily maximum predictions are a useful means of supplementing hourly predictions by capturing much of the upper variability.

4.6. Model validation

In contrast to the previous section, this section considers data that were not used in the model calibration or development process. The model was used to predict concentrations for 4 weeks at each site. These modelled concentrations are compared with measured values for the same time period. Table 2 shows the statistical parameters for the test data at each site for both 24 and 48 h hourly predictions and for daily maximum predictions. The FB indicates low bias for each of the time periods as expected as the MLR itself produces a model bias of zero. The FAC2 is sensitive to the magnitude of the data. It produces lower values at Kilkitt as concentrations are of magnitudes of a few ppb. At Winetavern Street where concentrations are highest, the FAC2 shows that 82% of the data satisfy the criteria for hourly values and 94% for daily maximum values. IA is the most robust test for model fit. Modelled results for Rathmines and Winetavern Street both have an IA of 0.90 for 24 h hourly predictions and 0.90 and 0.94, respectively for daily maximum predictions. The value is still high at Dun Laoghaire and Kilkitt (0.77 and 0.75, respectively for hourly values and 0.89 and 0.78 for daily maximum values) and but the overall model fit is thought to be lower due to the lower anthropogenic influence at

Table 2
Hourly average and daily maximum NO₂ predictions at each site (validation data).

		Winetavern Street		Rathmines		Dun Laoghaire		Kilkitt	
		24 h	48 h	24 h	48 h	24 h	48 h	24 hr	48 hr
Hourly	FB	−0.14	−0.04	−0.19	−0.21	0.07	0.07	−0.04	−0.09
	<i>R</i> (adj)	0.85	0.71	0.80	0.81	0.64	0.63	0.57	0.60
	FAC2	0.82	0.86	0.71	0.69	0.72	0.71	0.63	0.56
	IA	0.90	0.79	0.88	0.88	0.77	0.76	0.74	0.74
Daily maximum	FB	−0.13	−0.16	0.01	0.01	0.05	0.06	−0.05	−0.04
	<i>R</i> (adj)	0.92	0.89	0.84	0.83	0.79	0.77	0.54	0.56
	FAC2	0.94	0.88	0.96	0.96	0.90	0.94	0.71	0.71
	IA	0.94	0.92	0.90	0.90	0.89	0.88	0.78	0.71

Fig. 12. NO₂ concentrations for validation data at Winetavern Street hourly forecasts (top) and daily maximum forecasts (bottom).

these sites (less routine predictability).

Figs. 12–15 show the modelled versus monitored hourly and daily maximum NO₂ concentrations for the four test weeks at Winetavern Street, Rathmines, Dun Laoghaire and Kilkitt. Each

graph contains data for one week in spring, summer, autumn and winter.

At Dun Laoghaire the hourly model predicts the general increase in concentrations in spring but does not capture all of the short

Fig. 13. NO₂ concentrations for validation data at Rathmines hourly forecasts (top) and daily maximum forecasts (bottom).

Fig. 14. NO₂ concentrations for validation data at Dun Laoghaire hourly forecasts (top) and daily maximum forecasts (bottom).

term peaks and troughs on the 14th and 15th of March. The daily maximum model predicts these peaks with better accuracy. Both models predict the decrease towards the end of the week well. Overall predictions at Rathmines for this week are good (both the hourly and daily maximum models) with minimal smoothing of

peaks and troughs. A similar result is observed at Winetavern Street where the hourly model captures the general downward trend in concentration levels over the course of the week. The daily maximum model performs particularly well with one over prediction on the 16th of March. Concentrations at Kilkitt are low for

Fig. 15. NO₂ concentrations for validation data at Kilkitt hourly forecasts (top) and daily maximum forecasts (bottom).

this week and the hourly model predicts the mean concentration well whilst the daily maximum model provides an improvement on peak predictions.

During the summer and autumn weeks, the models perform particularly well at Rathmines and Winetavern Street. The hourly model captures short term fluctuations and the daily maximum model predicts most of the peak values which occur. The hourly model predicts the general increase in concentration levels at Dun Laoghaire well but misses some short term peaks towards the end of the first week. These are still under predicted by the daily maximum model but there is a significant improvement. Concentrations at Kilkitt for the first week were very low and the model predicts this well. It misses some peaks which are better captured by the daily maximum model. Concentrations are higher in the autumn test week and both models indicate this but due to the relatively small scale fluctuations which are involved neither are able to capture all variations.

The week in winter is well predicted by the model at all sites. There is some slight over prediction of maximum concentrations at Rathmines and Winetavern Street. The model performs well in predicting an unusually high peak at Kilkitt but is slightly out of phase with measurements. There is subsequently some over prediction as concentrations decrease. The utilisation of persistence within the model means that it can be slow to respond to sudden changes in concentration.

5. Discussion

This combined parametric and non-parametric modelling technique has been found to perform well in predicting both hourly NO₂ concentrations and daily maximum NO₂ concentrations out to 48 h. The hourly model predicts general concentration variation well at hourly resolution, indicated by high IA values. On occasion it can miss peak events (highlighted by reduced R values). While the hourly model tends to over smooth the data, the daily maximum model compliments it by capturing high concentration events while not impacting on average concentration predictions. This model achieves a high IA and FAC2 value. Combining these two models allows peak concentrations to be forecast and the short term variability in concentration levels to be predicted.

The model was found to perform better at the urban sites than at the rural site. There are a number of reasons for this. Firstly, concentrations at the rural site are very low, which means that the monitoring instrument is often not sufficiently precise to measure near zero concentrations. As a result these values are estimated to be equal to zero or the nearest 0.01 µg/m³, which leads to an unnatural distribution within the data. Therefore, while the modelled data follow the measured data reasonably closely, the statistical tests do not account for this lack in precision and indicate poorer results. Secondly, rural sites are less impacted by local anthropogenic activities which tend to be repetitive and cyclical (e.g. rush hour traffic). Since emissions travel a greater distance prior to reaching the rural monitoring site, there is more opportunity for dispersion and transformation of pollutants. While this results in lower NO₂ concentrations, it also leads to more variability in concentration levels (albeit at much lower total concentrations).

This model has a number of advantages over singular time series forecasting or the application of any of the individual methods in isolation. Use of a time series model alone would be unlikely to account for high or low events caused by external factors such as wind speed/direction or atmospheric stability, due to the smoothing effect resulting from the computation of seasonal indices. However, the ability of such models to describe cyclical effects such as transport emissions reduces the need to forecast such emissions directly. This is useful in particular when a detailed, spatially

resolved emission inventory is unavailable for the required area (as is the case in Ireland). The presented model requires no emissions data and the influence of emissions is instead extracted from historical monitored data. However, it should be noted that if major changes in the emission source were to occur (such as a sudden increase in road traffic volume), the model would require recalibration.

A typical approach to modelling is often to try a number of different models and select that with the most accurate result. However, the final selected model may not be the most applicable in the future due to changing parameters. Combining different methods (as is done in this paper) reduces this model selection error. It is also the case that relationships between variables are rarely pure linear or nonlinear and this combination of methods uses both linear and non-linear techniques decreasing the impact of the error due to any single assumption. Furthermore, in the event that certain data are not available for a given method (e.g. the meteorological forecast is not received on time), this model can still provide a prediction by making use of the time series forecast and the seasonal and diurnal factors.

This model benefits from simplicity of the input data and requires very low computational resources to run making it ideally suited to providing fast and reliable real time air quality forecasts.

6. Summary

A model that combines parametric and non-parametric modelling techniques to provide real time hourly forecasts of NO₂ has been presented. A large amount of variation in concentration levels at both rural and urban sites is explained by nonlinear relationships between wind speed, wind direction and concentration levels. A non-parametric technique, previously developed by the authors, has been applied to develop linear concentration factors for a given wind speed and direction pair and also to produce seasonal and diurnal factors. The model employs a multiple regression analysis driven by the above factors together with meteorological parameters. Model calibration and validation showed good agreement between measured and modelled data at predictions out to 48 h. The major advantages of this model include its low computational resources, easily available input data and minimisation of assumption based errors. Applied individually each technique is not sufficient to provide useful daily forecasts but the combination of parametric and non-parametric techniques allows a large degree of temporal variation in concentration levels to be explained. The model produces accurate forecasts and has been found to be a useful means of real time air quality forecasting at locations where monitoring data are available.

Acknowledgements

This paper has been written as part of the Science, Technology, Research and Innovation for the Environment (STRIVE) Programme 2007–2013 under grant number 2012_EH-FS-6. The programme is financed by the Irish Government under the National Development Plan 2007–2013 and administered on behalf of the Department of the Environment, Heritage and Local Government by the Environmental Protection Agency which has the statutory function of co-ordinating and promoting environmental research.

References

- Abdul-Wahab, S., Bakheit, C., Al-Alawi, S., 2005. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environ. Model. Softw.* 20 (10), 1263–1271.
- CEU, 2008. DIRECTIVE 2008/50/EC of the European Parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe. *E. Comm. Off. J.*

- Eur. Union L152/1.
- Chaloulakou, A., Saisana, M., Spyrellis, N., 2003. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Sci. Total Environ.* 313, 1–13.
- Cobourn, W., 2007. Accuracy and reliability of an automated air quality forecast system for ozone in seven Kentucky metropolitan areas. *Atmos. Environ.* 41, 5863–5875.
- Donnelly, A., Misstear, B., Broderick, B., 2011. Application of nonparametric regression methods to study the relationship between NO₂ concentrations and local wind direction and speed at background sites. *Sci. Total Environ.* 409 (6), 1134–1144.
- Donnelly, A., Broderick, B., Misstear, B., 2012. Relating background NO₂ concentrations in air to air mass history using non-parametric regression methods: application at two background sites in Ireland. *Environ. Model. Assess.* 17 (4), 363–373.
- European Environment Agency (EEA), 2011. The Application of Models Under the European Union's Air Quality Directive: a Technical Reference Guide.
- Elbir, T., 2003. Comparison of model predictions with the data of an urban air quality monitoring network in Izmir, Turkey. *Atmos. Environ.* 37, 2149–2157.
- Environmental Protection Agency, 2012. Ireland's Environment - an Assessment. Johnstown Castle.
- Gokhale, S., Raokhande, N., 2008. Performance evaluation of air quality models for predicting PM₁₀ and PM_{2.5} concentrations at urban traffic intersection during winter period. *Sci. Total Environ.* 394 (1), 9–24.
- Henry, R., Chang, Y., Spiegelman, C., 2002. Locating nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction. *Atmos. Environ.* 36 (13), 2237–2244.
- Jensen, S.S., 1998. Background Concentrations for Use in the Operational Street Pollution Model (OSPM). National Environmental Research Institute.
- Karppinen, A., Kukkonen, J., Elolahde, T., Konttinen, M., Koskentalo, T., 2000. A modelling system for predicting urban air pollution: comparison of model predictions with the data of an urban measurement network in Helsinki. *Atmos. Environ.* 34 (22), 3735–3743.
- Lissens, G., Mensik, C., Dumont, G., 2000. SMOGSTOP: a new way of forecasting ozone concentrations at ground level. *Int. J. Environ. Pollut.* 14 (1–6), 418–424.
- Martin, A., Barber, F., 1981. Sulphur dioxide, oxides of nitrogen and ozone measured continuously for 2 years at a rural site. *Atmos. Environ.* 15 (4), 567–578.
- Mavroidis, I., Iliä, M., 2012. Trends of NO_x, NO₂ and O₃ concentrations at three different types of air quality monitoring stations in Athens, Greece. *Atmos. Environ.* 63, 135–147.
- Mazzeo, N., Venegas, L., Choren, H., 2005. Analysis of NO, NO₂, O₃ and NO_x concentrations measured at a green area of Buenos Aires city during wintertime. *Atmos. Environ.* 39 (17), 3055–3068.
- O'Dwyer, M., 2012. Air Quality in Ireland 2011. Environmental Protection Agency, Johnstown Castle.
- Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC.
- Takezawa, K., 2005. Introduction to Nonparametric Regression. Wiley-Blackwell.
- Voukantsis, D., Karatzas, K., Kukkonen, J., Rasanen, T., Karppinen, A., Kolehmainen, M., 2011. Intercomparison of air quality data using principal component analysis, and forecasting of PM₁₀ and PM_{2.5} concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Sci. Total Environ.* 409 (7), 1266–1276.
- Willmott, Cort J., 1982. Some comments on the evaluation of model performance. *Bull. Amer. Meteor. Soc.* 63, 1309–1313.
- Yu, K., Cheung, Y., Cheung, T., Henry, R., 2004. Identifying the impact of large urban airports on local air quality by nonparametric regression. *Atmos. Environ.* 38 (27), 4501–4507.
- Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part I: history, techniques, and current status. *Atmos. Environ.* 60, 632–655.