An investigation into the association of ozone with traffic-related air pollutants using a quantile regression approach

S. Munir, H. Chen & K. Ropkins
Institute for Transport Studies, University of Leeds, UK

Abstract

Ground-level ozone (O_3) is one of the most harmful air pollutants due to its adverse effects on human health, agricultural crops, biodiversity and materials. Ozone is a secondary air pollutant and interacts with meteorological variables as well as with many other air pollutants such as nitric oxide (NO), nitrogen dioxide (NO_2), particles ($PM_{2.5}$), and carbon monoxide (CO). This paper intends to investigate the relationship of ozone with these air pollutants and lagged ozone (previous day ozone) at a roadside monitoring site in Leeds UK. A quantile regression approach has been applied, which is suitable for the non-normal ozone distribution and capable of handling nonlinearities in the associations of ozone with its predictors; as it examines the entire distribution of the variables rather than a single measure of central tendency (mean or median). Our results show that lagged ozone has positive, whereas NO, NO_2 and CO have negative associations with ozone. $PM_{2.5}$ is negatively correlated with ozone at lower quantiles (below 0.6) and the relationship becomes positive at upper quantiles (0.6 and above), perhaps indicating more complex interactions. Also, it is shown that the effect of explanatory variables on ozone concentrations is a function of quantiles and hence the behaviour and interaction of the covariates with ozone change at different regimes of ozone concentrations, information which is normally hidden in the traditional regression models. Further statistical analysis demonstrates that for some air pollutants the nature of relationship (negative or positive) between ozone and its predictors remains unchanged and only the strength changes, for others nature and strength both change at different quantiles. The study explores the impacts of traffic-related air pollutants on ground level ozone concentrations and suggests the use of quantile regression

WIT Transactions on Biomedicine and Health, Vol 15, © 2011 WIT Press
www.witpress.com, ISSN 1743-3525 (on-line)
doi:10.2495/EHR110031

approach for ozone and air quality data analysis as an alternative to traditional regression models.

Keywords: quantile regressions, ozone, air pollutants, NOx, CO, $PM_{2.5}$, lagged ozone.

1 Introduction

Background ozone concentrations over the last 20 years (1987 to 2007) have increased [1–3]. This increase in baseline concentrations is attributed to long distance migration of ozone from across the North Atlantic [2]. At the same time Jenkin [3] reported that local-scale removal of ozone by direct reaction with emitted NO has gradually decreased, a trend that is now widely attributed to on-going improvement in vehicle NOx emission regulations and associated progressive policy practices. This combination has resulted in a general increase in ozone concentrations since about 1990, which is most apparent at urban sites, but which to a less extent also influences the observations at the majority of rural locations. Air Quality Expert Group [2] has expressed their concerns that ozone levels in urban areas are increasing at comparatively faster rate than the surrounding rural areas, which in future may result in urban ozone levels as high as in the surrounding rural areas. If that happens it may increase ozone related health and environmental risks due to higher human exposure. Therefore it is vital to understand uncertainties in ozone predictions and quantify accurately the relationship of ozone with its sources and sinks.

Ozone is a regional pollutant and affects human health, agricultural crops, biodiversity and materials globally and exhibits distinct regional trends. Ozone concentrations also vary spatially from place to place within the UK considerably. Roadsides, urban centres, rural areas and remote sites all show different characteristics in terms of ground-level ozone. Ozone concentration at a given location is not only dependent on meteorological variables, but also on the concentrations of other air pollutants, e.g. NOx, CO, hydrocarbon etc. Several scientists have investigated the relationship of ozone with different air pollutants (e.g. [2, 4, 5]) and have reported that these air pollutants play a vital role in ozone formation (e.g. ozone precursors i.e. NOx and HC) and destruction (e.g. NO). Therefore, for accurate ozone prediction it is important to understand their mutual interaction and the role they play in controlling ozone concentrations.

Different techniques (models) have been used to study ozone and its associations with meteorological factors as well as with other air pollutants. Models have been used to predict ozone concentrations, establish long or short term ozone trends, understand underlying mechanisms in the formation and destruction of ozone, and study the health and environmental impacts of ozone [6, 7]. Multiple linear regressions are the most widely used methodologies for modelling the dependence of ozone on several independent variables (predictors). Soja and Soja [8], Tidblad et al. [4], Paschalidou et al. [5] and Pont and Fanton [9] all applied multiple regressions for ozone modelling. Linear regressions explicitly assume normality and linearity of the data, which are not

met by ozone and other air pollutants data. This study uses a quantile regression approach that is applicable to both normal and non-normal distributions and is capable of handling the non-linearities in ozone and other air pollutants data. Quantile regression model is especially useful when extremes values are important, such as air quality studies where upper quantiles of air pollutant (e.g. ozone) levels are critical from a public health perspective.

2 Methodology

This study is mainly based on the statistical analysis of ozone, NOx, CO, lagged ozone, and $PM_{2.5}$ data measured at Kirkstall roadside monitoring site in Leeds UK for a 2 year period. The data is divided into two subsets: training (Nov 2007 to Oct 2009, except May 2009) and test dataset (May, 2009). The study is applying a quantile regression approach, which has been explained in section 2.2.

2.1 Monitoring sites

Most of the data used in this study are taken from Kirkstall roadside monitoring site, which is part of the facilities available at Institute for Transport Studies (ITS) University of Leeds for the monitoring of air pollution, traffic and meteorological variables. The monitoring station lies between 53°48'31.38"N and 1°35'21.40"W, with Kirkstall Road (A65) running North-West to South-East adjacent to the site. The road is a busy thoroughfare with nearby petrol pump (20 meters) and used car garage (50 meters) to the South. In addition to ozone, the site monitors CO, NO_X and hydrocarbons (HCs) using certified gaseous analysers. This site also has facilities for monitoring wind speed, wind direction, temperature, humidity and solar radiation. Data from Harwell air quality monitoring station have also been analysed, which is a rural site and is part of the UK AURN (automatic urban and rural network), see [10] for the details of AURN sites.

At ITS every effort is made to ensure the quality of data is maintained. Automatic nightly calibrations of gaseous analysers, and fortnightly 'manual' zero and span calibrations using calibration gases CO, NO, NO_2 and Benzene are performed routinely. After collection the data go through verification, a process to clean-up the initial data. The data from AURN go through a proper 'data verification and ratification process' before it is marked as 'Ratified' data. All the data from AURN have a standard Quality Assurance and Quality Control (QA/QC).

2.2 Quantile regression model

This study applies quantile regression model (QRM) proposed by Baur et al. [6] for ozone and air quality data analysis and has certain advantages over other methods. QRM can be used for both parametric and nonparametric regression methods, as this model does not depend on the single measure of the central tendency (mean or median) of the data distribution only; instead it examines the

entire distribution of the data and hence is robust to departures of the data from normality and skewed tails. QRM allows the covariates to have different impacts at different points of the data distribution and is, therefore, capable of handling the non-linearities in the association of dependent and independent variables.

The linear regression model (LRM) focuses on modelling the conditional mean of a response variable (in our case ozone) without addressing its full distribution, whereas the quantile regression model accommodates analysis of the full distribution of the response variable. The QRM estimates the potential differential effect on various quantiles of the data distribution. In general form the QRM is presented as below [11]:

$$\mathrm{yi} = \beta_0^{(\mathrm{p})} + \beta_1^{(\mathrm{p})}\mathrm{xi} + \varepsilon_\mathrm{i}^{(\mathrm{p})} \tag{1}$$

$$\mathrm{yi} = \beta_0^{(\mathrm{p})} + \sum_{\mathrm{i}=1}^{\mathrm{k}} \beta_\mathrm{k}^{(\mathrm{p})}\,\mathrm{xik} + \varepsilon_\mathrm{i}^{(\mathrm{p})} \tag{2}$$

where p shows the pth quantile and $0 < p < 1$, y represent the response variable, x the explanatory variable, β_0 (constant) the intercept, β_1 the slope (gradient) and ε the error term. ε the error term in LRM is assumed to be independent of the value of the covariates (homoscedasticity). In contrast, quantile regression models allow for the variance of the error term to vary (heteroscedasticity) and make no assumptions about the variance structure. Moreover, the pth quantile of the error term conditional on the regressor is assumed to be zero i.e. $\varepsilon_\mathrm{i}^{(\mathrm{p})} = 0$, which make equation 2 as:

$$\mathrm{Qyi}\ (\mathrm{p}|\mathrm{xi1} \ldots \mathrm{xik}) = \beta_0^{(\mathrm{p})} + \sum_{\mathrm{i}=1}^{\mathrm{k}} \beta_\mathrm{k}^{(\mathrm{p})}\,\mathrm{xik}.$$

The constant β_0 and the coefficients β_1 are estimated for 99 different quantiles (p=0.01,…, 0.99) using each time the entire dataset. The 0.5^{th} quantile represent the median, half of the data occur above the median and half below the median.

R (2.12.0) and two additional packages 'openair' and 'quantreg' were used to perform the statistical analysis presented in this report.

3 Results and discussion

The distributions of ozone and the other air pollutants were studied and it was established that their distributions were non-normal, and therefore Spearman correlation was applied which is a distribution free method for finding the correlation between two variables. Ozone concentrations have been shown to have strong correlation with these covariates. The Spearman correlation coefficients (R) for hourly mean data between ozone and other air pollutants were -0.64, -0.70, -0.68, -0.51, -0.53, 0.47 for NO, NOx, NO_2, CO, $PM_{2.5}$ and lagged-ozone respectively. The correlation between ozone and other variables was negative, except lagged-ozone which showed a positive correlation with ozone.

The outputs of quantile regression model are shown in Figure 1, using ozone as a response variable and lagged-ozone, NO, NO_2, CO and $PM_{2.5}$ as explanatory

variables. The Barrodale and Roberts (br) algorithm method for computing the fit has been adopted here. The 'br' method has been described in details in Koenker and d'Orey [12] as an efficient technique for large datasets (e.g. up to several thousand observations). In Figure 1 alongside quantile regression, the outputs of ordinary least square regression have also been visualised. In ordinary least square regression, only one regression coefficient represents the entire distribution of the explanatory variable (indicated by solid line along with its 95% confident interval); whereas in quantile regression generally several coefficients are given depending on the number of quantiles chosen. In Figure 1 regression coefficients have been given for 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99 quantiles (represented by dashed-dotted line with their 95% confident intervals). In Figure 1 the top left panel shows the intercepts of the model. The values of intercept (constant) are higher for higher quantiles and vice versa. For instance, the intercept value for 0.1 quantile is about 17; whereas it is about 40 for quantile 0.99. Detailed analysis of the model outputs is described in the following sections.

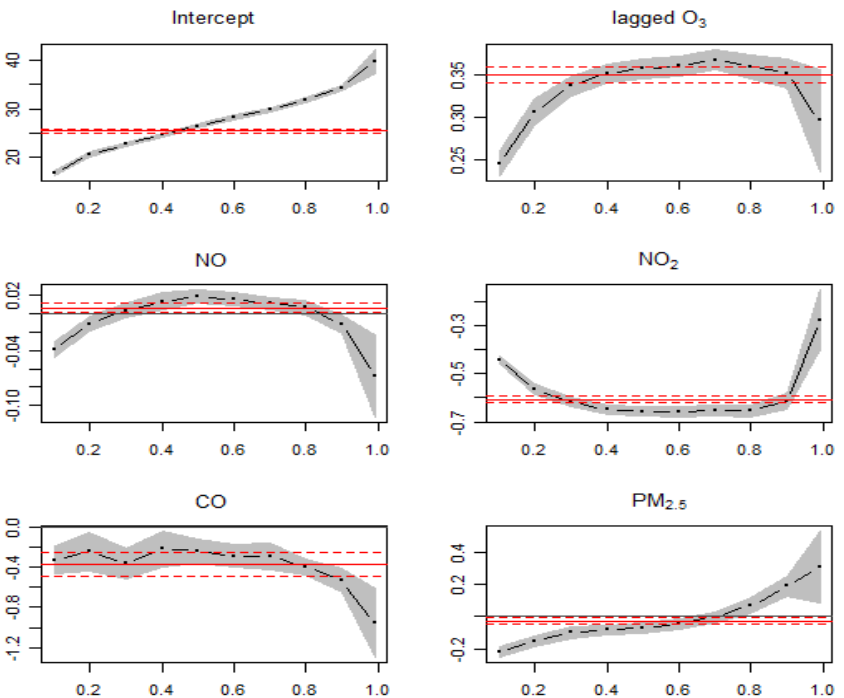

Figure 1: The outputs of quantile regression model showing the effect of lagged-ozone, NO, NO_2, CO and $PM_{2.5}$ on hourly mean ozone concentrations. Quantile regression coefficients (dashed-dotted line) and ordinary least square regression coefficients (solid line) are presented with 95% confidence interval. Various quantiles are shown on x-axis, whereas their coefficients are shown on y-axis.

3.1 Auto-regression analysis ozone vs. lagged ozone

Lagged ozone (previous-day hourly mean ozone ppb) has positive effect on ozone mixing ratios. Figure 2 shows a scatter plot between ozone and lagged-ozone data from Kirkstall site for May, 2008. Ten estimated quantile regression lines for different values of Quantiles (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99) have been superimposed on the scatter plot. The median (0.5 quantile) is indicated by bold broken line and the least squares estimate of the conditional mean function by bold solid line. There is a clear positive correlation between ozone and lagged-ozone, i.e. increasing lagged-ozone results in increasing ozone mixing ratios. The effect of lagged-ozone varies with quantile, as depicted in Figure 1 (top, right). The strength of relationship increases with increasing quantile values until quantile 0.7 and decreases afterward. At higher quantiles the lower coefficient values shows low persistence of ozone at extreme concentrations. The confidence bands are wider at higher quantile (0.99) showing less accurate modelling at these concentrations. On the other hand ordinary least square regression gives only one regression coefficient, which is represented by a straight line, as it considers only the mean value of the data and therefore hides the rest of the details.

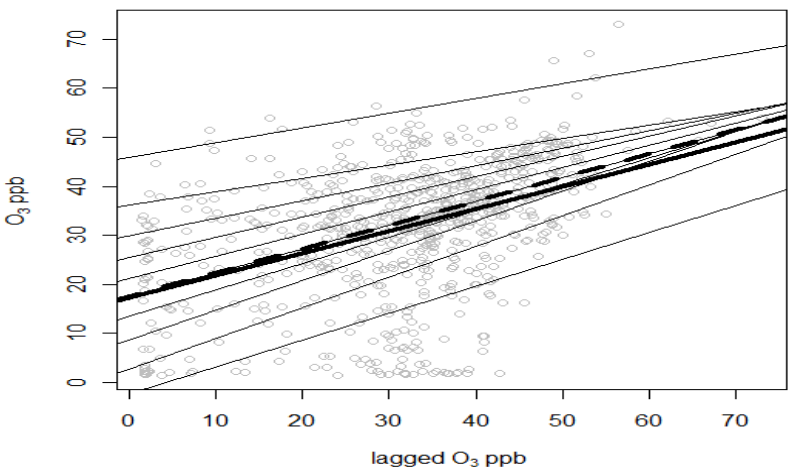

Figure 2: Scatter plot of ozone vs. lagged (previous day) ozone hourly mean data (May, 2008 from Kirkstall site in Leeds). Ten estimated quantile regression lines for different values of quantiles (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99) have been superimposed on the scatter plot. The median (quantile 0.5) is indicated by bold broken line; the least squares estimate of the conditional mean function is indicated by bold solid line.

3.2 Ozone and nitrogen oxides

NO and NO_2 are collectively known as NOx because they are rapidly inter-converted during the day. NO and NO_2 are both generated by combustion processes in the atmosphere, which mainly produce NO with a small proportion of NO_2 (~ 5%) [13]. Most of NO_2 is formed in the atmosphere by oxidation of NO, for example, by reaction with ozone. Therefore NO_2 is considered as a secondary (formed in the atmosphere) and NO as a primary pollutant (directly emitted). In the UK over 50% nitrogen oxides are produced by transport. NO_2 is split up by UV light to give NO and an oxygen (O) atom, which combines with molecular oxygen (O_2) to make ozone. In rural air, away from sources of NO, most of the nitrogen oxides in the atmosphere are in the form of NO_2, whereas near a source (e.g. a busy road) NO is the dominant species. Figure 3 shows the ratios of NO and NO_2 (NO/NO_2) at both Kirkstall (roadside) and Harwell (rural) monitoring site and confirms that the level of NO is more than NO_2 at the roadside monitoring site, whilst the opposite is true for the rural site. The reason is clear that at roadside traffic vehicles produce NOx which is mostly consists of NO and by the time these gases reach rural areas most of the NO is oxidised into NO_2.

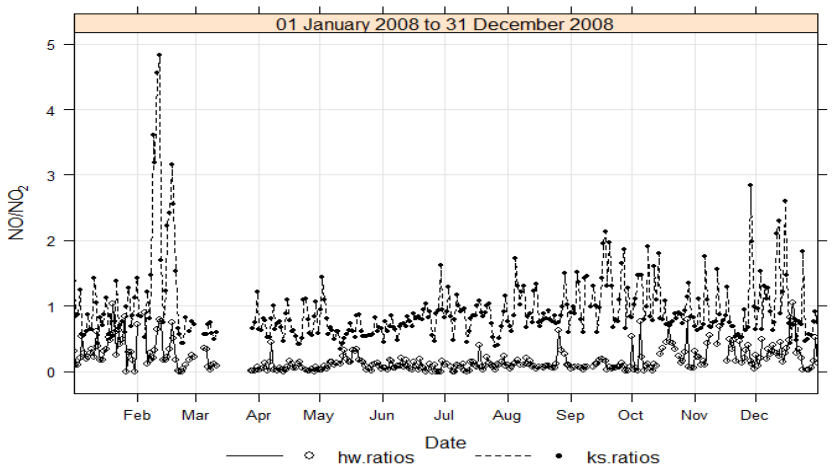

Figure 3: NO/NO_2 ratios for the year 2008 at Kirkstall (ks) roadside and Harwel (hw) rural site.

Quantile regression exhibits considerably stronger effect of NO_2 than NO on ozone mixing ratios. The quantile regression coefficients range from about '-0.06 to +0.02' for NO and '-0.3 to -0.7' for NO_2. It can be clearly seen in Figure 1 that the strengths of coefficients for NO and NO_2 follow opposite trends, i.e. for NO the highest correlation coefficients (absolute values) are observed at quantiles 0.1 and 0.99, whereas for NO_2 the weakest coefficients were recorded for these two quantiles. In other words, NO shows maximum effect whereas NO_2

shows minimal effect on ozone concentrations at extreme values (minimum and maximum). When ozone is modelled using only NO or NO_2 as explanatory variables with exactly the same quantiles, NO present a different picture (making a bowl shape as NO_2 does in Figure 1) , whereas NO_2 behave almost in the same way (Figure not shown here).

3.3 Ozone vs. CO and $PM_{2.5}$

In this section the association of ozone with CO (Figure 1, bottom-left) and $PM_{2.5}$ (Figure 1, bottom-right) is investigated using quantile regression model. CO has negative effect on ozone mixing ratios and the effect becomes stronger at quantile 0.9 and 0.99. The CO effect on ozone at different quantiles is not significantly different from the mean effect (as the confident intervals overlaps), except at 0.9 quantile and above. The effect of $PM_{2.5}$ on ozone mixing ratios is negative below 0.6 quantile and positive above. The magnitude of estimated coefficients (absolute value) of $PM_{2.5}$ decreases gradually from 0.1 to 0.6 quantile and become positive above 0.6 quantile. The effect gradually increases and reached a maximum value at 0.99 quantile. The negative coefficients of CO are most probably due to the fact that the data come from a roadside monitoring site and therefore almost all of CO is emitted by road traffic. Higher mixing ratios of CO pollutants indicate higher traffic volume and hence higher NO which depletes ozone.

4 Goodness of fit for quantile regression

The goodness of fit in ordinary least square regression is measured by the coefficient of determination (R^2), which is based on least squares criterion. R-squared values range from 0 to 1. Larger value of R-squared indicates a better model fit. In quantile regression the goodness of fit is represented by $R^1(\tau)$ and its values, like R^2, lies between 0 and 1[14]. R^2 measures a global goodness of fit over the entire conditional distribution, whereas $R^1(\tau)$ measures the local performance of model for a given quantile. Koenker and Machado [14] suggest measuring $R^1(\tau)$ by comparing the sum of weighted distance for the model of interest with the sum in which only the intercept is used (for details see [11] and [14]). $R^1(\tau)$ and R^2 have different nature, as the former is a local whereas the latter is a global measure of performance and therefore are not directly comparables. R^1 (τ) values for different quantiles have been shown in Figure 4, which are relatively weaker as compared to global goodness of fit.

To estimate a global goodness of fit (R^1) for quantile regression model, this study adopts the approach suggested by Baur et al. [6] and is called amalgated quantile regression model (AQRM). AQRM approach for estimating the performance of the model is simple and can be directly compared with R^2 for the linear regression. To estimate R^1, firstly quantile regression coefficients were determined for 10 quantiles (.1, .2, .3, .4, .5, .6, .7, .8, .9, .99) using ozone as variate and NO, NO_2, lagged-O_3, CO and $PM_{2.5}$ as covariates for the whole dataset. The test dataset (May, 2009) was divided into 10 equal subsets

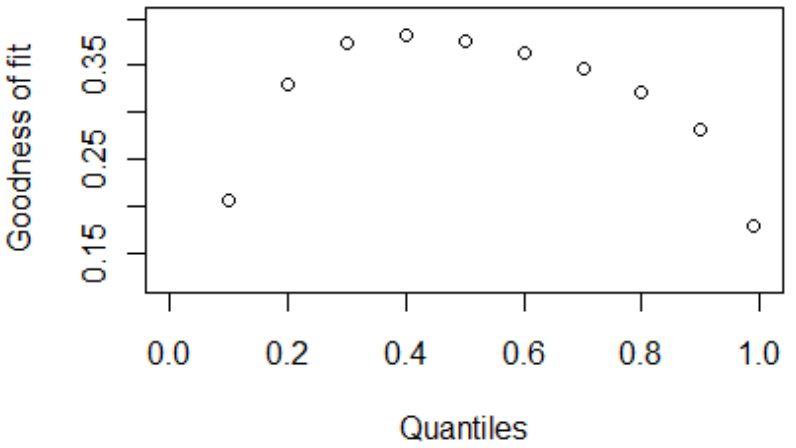

Figure 4: Local goodness of fit $R^1(\tau)$ as a function of the ozone quantiles for the quantile regression model at Kirkstall site Leeds.

according to the above quantile values of ozone data. Using quantile regression coefficients of each quantile, ozone was predicted for each subset. For the estimation of quantile regression coefficients the whole dataset was used as training data, except May 2009, which was used as test data for prediction purposes. Finally predicted and observed ozone were compared for the test data (Figure 5 and Figure 6).

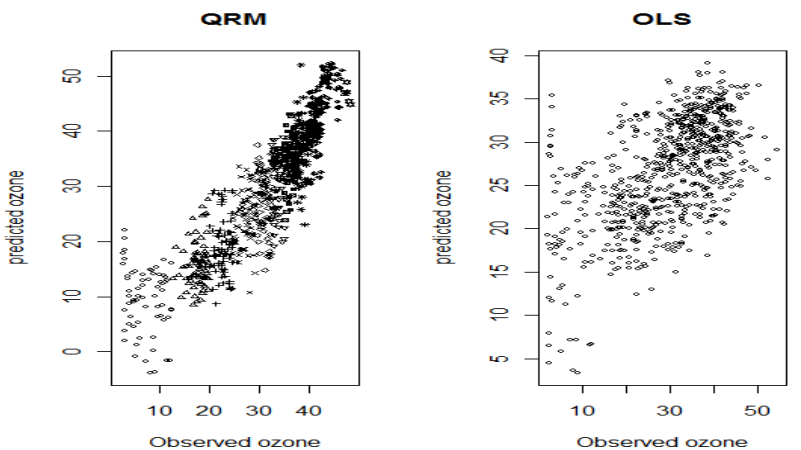

Figure 5: Predicted versus observed ozone concentration at Kirkstall site using AQRM (amalgated quantile regression mode) R^1=0.80) (left) and OLS (ordinary least square) model, R^2=0.53 (right) for May 2009.

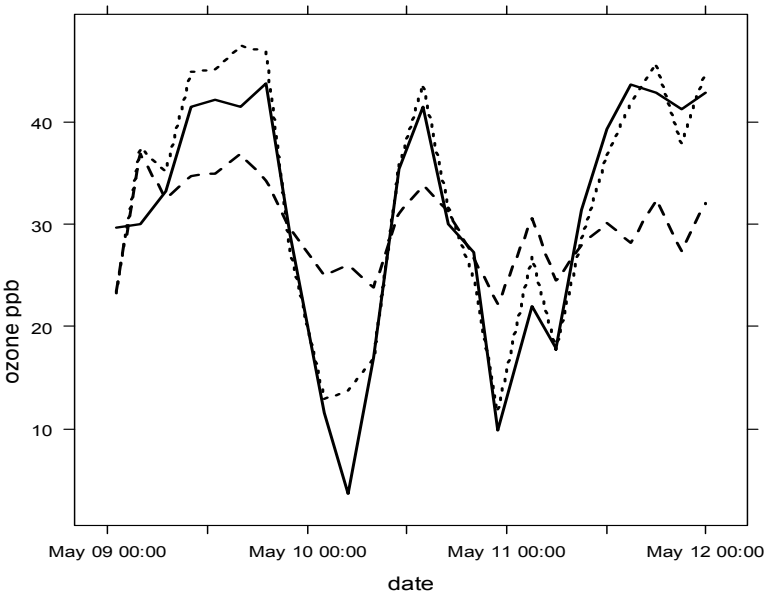

Figure 6: Observed (solid line) vs. predicted ozone, using AQRM (R^1=0.80, dotted line) and OLS model (R^2=0.53, dashed line) for 9 to 11 May 2009 at Kirkstall site.

Figure 5 depicts predicted ozone versus observed ozone mixing ratios at Kirkstall site. The scatter plot of observed ozone versus predicted ozone by Quantile Regression Model (QRM) is shown in the left, whereas the scatter plot of observed ozone versus predicted ozone by Ordinary Least Square model (OLS) is shown in the right panel of Figure 5. AQRM explains more of the ozone variations showing R^1-vlaue of 0.80 in comparison to OLS which gives R^2-value of 0.53. This indicates that AQRM is explaining significantly more ozone variation than OLS. QRM model was more efficient in predicting ozone mixing ratios than OLS model, particularly at extreme values as shown in Figure 6, where the dotted line (QRM) closely follows the line of observed ozone.

5 Conclusion

This study explores the impacts of traffic-related air pollutants (NO, NO_2, CO, $PM_{2.5}$) and lagged ozone on ground level ozone and suggests the use of quantile regression approach for ozone and air quality data analysis as an alternative to traditional regression models. Quantile regression model is suitable for non-normal ozone distribution and is capable of handling nonlinearities in the associations of ozone with its predictors; as it examines the entire distribution of the variables rather than a single measure of central tendency (mean or median).

It is shown that the effect of explanatory variables on ozone mixing ratios is better explained by quantiles and hence the behaviour and interaction of the variables with ozone changes at different regimes of ozone concentrations, which is normally hidden in the traditional regression models. Statistical analysis demonstrates that for some air pollutants the nature of relationship (negative or positive) between ozone and its predictors remains unchanged and only the strength changes, for others nature and strength both change at different quantiles, possibly indicating more complex interactions. Quantile regression model explains significantly more variations in ozone (R^1= 0.80) as compared to ordinary least square regression (R^2=0.53) and is therefore better suited for ozone data analysis and prediction.

Acknowledgement

We gratefully acknowledge Economic and Social Research Council (ESRC) for providing funding for this study, which is a part of my PhD project.

References

[1] Derwent, R.G., Simmonds, P.G., Manning, A.J. and Spain, T.G., Trends over a 20-year period from 1987 to 2007 in surface ozone at the atmospheric research station, Mace Head, Ireland, *Atmospheric Environment,* 41, pp. 9091–9098, 2007.

[2] Air Quality Expert Group (AQEG). Ozone in the UK, the fifth report produced by air quality expert group (AQEG), 2009. DEFRA publication London, 2009AQEG, 2009

[3] Jenkin, M.E., Utembe, S.R. and Derwent, R.G., Modelling the impact of elevated primary NO_2 and HONO emissions on regional scale oxidant formation in the UK, *Atmospheric Environment,*42 (2), pp. 323–336, 2008.

[4] Tidblad, J., Mikhailov, A.A., Henriksen, J. and Kucera, V., Improved Prediction of Ozone Levels in Urban and Rural Atmospheres, 40 (1), pp. 67–76, 2002.

[5] Paschalidou, A.K., Kassomenos, P.A. and Bartzokas, A., A comparative study on various statistical techniques predicting ozone concentrations: implications to environmental management, *Environmental Monitoring and Assessment*, 148 (1-4), pp. 277–289, 2008.

[6] Baur, D., Saisana, M. and Schulze, Modelling the effects of meteorological variables on ozone concentration-a quantile regression approach, *Atmospheric Environment*, 38 (28), pp. 4689–4699, 2004.

[7] Gardener, M.W. and Dorling, S.R., Meteorologically adjusted trends in the UK daily maximum surface ozone concentrations, *Atmospheric Environment,* 34, pp. 171–176, 2000

[8] Soja, G., and Soja, A.M., Ozone indices based on simple meteorological parameters: Potentials and limitations of regression and neural network models, *Atmospheric Environment*, **33**, pp. 4299–4307, 1999.

[9] Pont, V. and J. Fontan, Comparison between weekend and weekday ozone concentration in large cities in France, Atmospheric Environment, 35, pp. 1527–1535, 2000.
[10] UK automatic urban and rural network (AURN). Department for Environment, Food and Rural Affairs. www.aurn.defra.gov.uk. Accessed October 12, 2010
[11] Hao, L., Naiman, D.Q., Quantile regression: Series-Quantitative applications in the social sciences, Sage Publications, 2007 (Series NO. 07-149)
[12] Koenker, R.W. and d'Orey, Computing regression quantiles, *Applied Statistics*, 36, pp. 383–393, 1994
[13] Air Quality Expert Group (AQEG). Trends in primary nitrogen dioxide in the UK, the fourth report prepared by the air quality expert group, DEFRA Publication London, 2007.
[14] Koenker, R. and J. Machado, Goodness of fit and related inference processes for quantile regression, *American Statistical Association*, 94, pp. 1296–1310, 1999