PERGAMON

Atmospheric Environment 34 (2000) 21–34

ATMOSPHERIC ENVIRONMENT

www.elsevier.com/locate/atmosenv

Statistical surface ozone models: an improved methodology to account for non-linear behaviour

M.W. Gardner, S.R. Dorling*

School of Environmental Sciences, University of East Anglia, Norwich, UK

Received 26 March 1999; received in revised form 26 July 1999; accepted 26 July 1999

Abstract

Using UK data as a case study, this paper demonstrates that statistical models of hourly surface ozone concentrations require interactions and non-linear relationships between predictor variables in order to accurately capture the ozone behaviour. Comparisons between linear regression, regression tree and multilayer perceptron neural network models of hourly surface ozone concentrations quantify these effects. Although multilayer perceptron models are shown to more accurately capture the underlying relationship between both the meteorological and temporal predictor variables and hourly ozone concentrations, the regression tree models are seen to be more readily physically interpretable. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Artificial neural networks; CART; Decision trees; Air quality modelling

1. Introduction

The current understanding of the relationship between meteorology and surface ozone concentrations is well documented (Davies et al., 1992; O'Hare and Wilby, 1995; PORG, 1997; Kunz and Speth, 1997; Entwistle et al., 1997; Derwent et al., 1998) and attempts to develop statistical regression models of daily maximum ozone have been numerous (for example Feister and Balzer, 1991; Bloomfield et al., 1996; Hubbard and Cobourn, 1998). The development of these models involved the detailed choice of both site specific nonlinear data transformations and interactions between predictor variables in order to better capture the ozone behaviour. When developing models of hourly ozone concentrations, such model selection may become problematic due to the requirement for potentially complex nonlinear interactions between predictor variables. The primary objective of this paper is to quantify the importance of nonlinearities and interactions between predictor variables when developing models of hourly surface ozone. A secondary objective is to describe a practical methodology for the development of improved statistical models.

Recently, artificial neural networks, and in particular the multilayer perceptron (MLP), have found application within the atmospheric sciences (Gardner and Dorling, 1998). MLP models are described later in this paper, however one important feature is their ability to represent any smooth functional relationship between one or more predictor and predictand variables (Hornik et al., 1989). Hence any form of regression model, including those with nonlinear data transformations and arbitrary interactions between predictor variables, can be considered as a special case of MLP model. If the performance of a properly trained MLP model is similar to that obtained by a linear regression model then it can be confidently assumed that the functional relationship between the *chosen* input and output variables is linear. If MLP models outperform linear models then some form of nonlinearity or interaction between the input variables is required in order to represent the function being modelled. Comparison between linear and MLP models can therefore be used to identify the complexity of functional relationships. *Well-trained* MLP model

* Corresponding author. Fax: 01603-507719.

E-mail address: s.dorling@uea.ac.uk (S.R. Dorling)

PII: S1352-2310(99)00359-3

performance should always match or outperform a linear model developed on the same data.

Yi and Prybutok (1996) developed neural network models for predicting daily maximum surface ozone concentrations in Central Europe from station meteorological predictors. The neural network model performance was seen to be better than a regression model. However Comrie (1997) compared neural network and linear regression models of daily maximum ozone, in U.S. urban areas using meteorological predictors, and found only slight improvements when using neural network models. As noted in Gardner and Dorling (1998), this indicates that the relationship between the meteorological variables analysed and the daily maximum ozone concentration (at the sites considered) can be well represented by a linear model. Given additional predictor variables it is possible that this may no longer be the case. In this paper comparisons between MLP and linear models are made to quantify the complexity of the ozone–meteorology relationship at the hourly time scale at a number of U.K. urban and rural sites. In addition regression tree models are also applied to the same data. Regression trees (which are more fully described later) have been used to model daily maximum ozone concentrations in Canada (Burrows et al., 1995). Whilst capable of capturing some nonlinearities and interactions between predictor variables they are not quite as expressive as MLP models, in the sense that they cannot easily represent as broad a range of functional mappings. However regression trees, unlike MLP models, are fully interpretable.

2. MLP Neural networks

Gardner and Dorling (1998) describe the practical application of MLP neural networks, hence only important issues will be raised here. As stated in the introduction, the main advantage provided by MLP neural networks is their ability to represent any smooth measurable functional relationship between one or more predictor and predictand variables. The nature of the functional relationship is learnt during a supervised training process directly from data. For many problems, identifying the nature of the functional mapping is difficult and in such situations MLP models are particularly useful. Inappropriate design of regression models frequently results in MLP models demonstrating superior performance when developed on the same data.

MLP models could be considered as over parameterised; they are capable of memorising the data upon which they are developed. Such a model is described as overtrained and will generalise poorly when tested on unseen data. To avoid overtraining, and to restrict model complexity, a regularisation technique known as early stopping can be used (Sarle, 1995). This involves splitting the available data into a training, validation and test set. The MLP model is trained using the training data. During training the validation set is used to gauge the generalisation performance of the model. At the point when overtraining occurs the generalisation performance will start to decrease. Training should be stopped at this point and then the performance of the trained MLP model can be assessed using the third independant test data set. Alternative regularisation techniques are available (see for example Bishop, 1995) however early stopping is generally favoured, and was used here, due to its' simplicity.

The MLP models described in this paper were trained using the scaled conjugate gradient (SCG) algorithm (Moller, 1991). Training involves finding the set of MLP network weights which enable the MLP model to represent the underlying patterns in the training data. This is achieved by minimising the MLP network error, for all input patterns, with respect to the associated network output patterns. Unfortunately the error surface is often complex and contains many local minima (Comrie, 1997). Should the training algorithm become trapped in a local minimum the final MLP model may be suboptimal. Typically the global minimum is not reached and a good local minima is treated as an acceptable solution. Training a number of MLP models and selecting the model with the best generalisation performance can reduce the likelihood of local minima causing problems. The SCG algorithm has been shown to be superior to standard backpropagation and is also faster than other conjugate gradient techniques (Moller, 1993). Unlike alternative algorithms, the scaled conjugate gradient algorithm is insensitive to the choice of initial training parameters and has a high probability of converging to a good solution. The Stuttgart Neural Network Simulator (freely available via the Internet from ftp://ftp.informatik.uni-stuttgart.de) was used. This simulator runs under UNIX and enables MLP models to be easily designed and trained.

In the same manner as was reported in Gardner and Dorling (1999a), large MLP models with two hidden layers were used. Although theoretically only one hidden layer is required, two hidden layers can speed up convergence and reduce training time. Networks with two hidden layers are more likely to escape poor local minima during the training process. The use of large networks to speed up training requires that care is taken to avoid overtraining (Sarle, 1995). For all MLP models, the transfer function in the hidden layer nodes was the hyperbolic tangent function whilst for the input and output layer nodes the unbounded identity function was used.

3. Regression trees

MLP models can be constructed that represent functional mappings with a high degree of accuracy. It would be desirable to discover exactly what the models have learnt in order to more fully understand the system being modelled. Unfortunately at present no methods exist to analyse the weights of a trained MLP model and to deduce simple physical explanations of the model behaviour.

Regression trees provide an alternative statistical modelling technique which, when compared to the MLP, are much more interpretable models. CART (Classification and Regression Trees) is described by Breiman et al. (1984) and represents one type of tree induction algorithm, alternatives are assessed by Lim et al. (1997). CART can be described as a nonparametric-data-driven rule generating algorithm; nonparametric in the sense that the number of parameters in the final model is not specified beforehand. Classification trees are used when the predictand takes a discrete number of values whereas regression trees are used when the predictand is continuous. Regression trees can represent non-linear functions in a piecewise, discretised manner. The output from a CART analysis consists of a binary tree structure and associated with each split in the tree is a rule involving one or more predictor variables. In addition to their transparency, tree based models are quicker to construct than MLP models and during development irrelevant predictor variables are removed from the analysis automatically.

The tree is generated via a recursive partitioning algorithm which determines the rules associated with each binary split in the tree. Initially all of the data are used to determine the rule which splits the data into the two most dissimilar sets. A rule consists of a variable and a threshold, for example temperature less than 15°C. Following the initial split, each subset is recursively split until a large tree is developed with each terminal node containing only a few observations. The oversized tree is then pruned until the generalisation performance, either measured directly with independant data or estimated using cross-validation, indicates no overfitting. The final tree can then be used with new data by traversing the tree with the new data vector. In this work the CART algorithms that are included in the S-Plus statistical software package were used, as described by Venables and Ripley (1994).

4. Data

Hourly ozone data were obtained from the UK Department of the Environment, Transport and the Regions (DETR) automatic air quality monitoring network (http://www.aeat.co.uk/netcen/aqarchive/). This was matched with UK Meteorological Office (UKMO) hourly synoptic meteorological data hosted by the British Atmospheric Data Centre (BADC, http://www.badc.rl.ac.uk/). Suitable sites were selected based on the availability of important meteorological variables and the proximity of the ozone and meteorological stations. It was also desirable that the data could be divided into independant training, validation and test data sets. For this reason a minimum of four years of data coverage was required to allow two years of data for training and one year for both validation and testing purposes.

Based on earlier work (Gardner and Dorling, 1996) the meteorological variables considered were the amount of low cloud, base of lowest cloud, visibility, dry bulb temperature, vapour pressure, wind speed and direction. The choice of meteorological variables was also based upon selecting those which were generally available and routinely forecast, since this would facilitate the results of this work being applied in a forecasting mode. Ideally solar radiation data would have been used since this is a good indicator of the likelihood of photochemical ozone production. Unfortunately, hourly radiation data were not available for the majority of meteorological stations, however it was expected that the combinations of cloud cover, temperature and visibility would together act as reasonable surrogates for solar radiation.

Five sites satisfied all criteria and these are listed in Table 1 along with the distance between the meteorological and air quality monitoring sites. For all sites except Southampton data were available for the 4 yr period 1993–1996. At Southampton a slightly incomplete data set was used covering the period 1994 to 1997. It was

Table 1
DETR automatic network ozone monitoring sites and the UKMO meteorological sites used, and the approximate distance between them

DETR station name	DETR station type	UKMO station name	Distance (km)
Bristol centre	Urban central	Bristol weather centre	1.9
Edinburgh centre	Urban central	Turnhouse	9.7
Eskdalemuir	Rural	Eskdalemuir	1.4
Leeds centre	Urban central	Leeds weather centre	0.5
Southampton centre	Urban central	Southampton weather centre	2.4

unfortunate that only one rural site (Eskdalemuir) emerged as being suitable since this introduces an urban bias to the results.

5. Methodology

MLP, regression tree and linear regression models of hourly surface ozone were developed using the hourly meteorological predictor variables. In addition to the meteorological data, models with additional seasonal inputs were also developed. Gardner and Dorling (1996) demonstrated that seasonal inputs, consisting of $\sin(2\pi d/365)$ and $\cos(2\pi d/365)$ where d is the Julian day of the year, enabled MLP models to more accurately resolve the seasonal signal in ozone concentrations. It was also envisaged that time of day may be a useful predictor in urban environments where, for example, vehicular NO_x emissions, which are strongly related to time of day, could influence ozone concentrations (Derwent and Davies, 1994). The models, including the MLP model architectures, are described in Table 2.

The MLP, regression tree and linear models were calibrated using the training data set. The validation data set was used for early stopping when training the MLP models, and for pruning the regression tree models. All models were finally assessed using the test data set. The linear models were developed in the most simplistic manner without data transformations or variable interactions.

6. Results

Table 3 presents the performance statistics for all the models when tested on the independant test data. A range of statistics, described below, are presented. It was noted by Fox (1981) and subsequently by Willmott (1982) that inconsistencies exist in the manner with which air quality models, and models more generally, are tested and evaluated. It is not uncommon to find models described in the literature with performance assessed by nothing more informative than a correlation coefficient. As noted by Comrie (1997) such statistics on their own provide no useful information concerning the accuracy of model predictions, and seriously hinder model intercomparisons.

Mean bias error (MBE), defined as the difference between the mean predicted and observed concentrations, indicates the degree that observed concentrations are over or underpredicted. The observed (s_o) and modelled (s_p) concentration standard deviations quantify the amount of the variance the model is capturing compared to the observed data. Two commonly reported measures of residual error include the mean absolute error (MAE) and the root mean squared error (RMSE) which summarize the difference between the observed and modelled concentrations. Due to the power term in the RMSE calculation it is more sensitive to extreme values than MAE.

Although correlation measures do not give any indication of model accuracy they are also commonly reported. The coefficient of determination (R^2) is an intuitively attractive statistic since it indicates how much of the variability in the observed data is being reproduced by the model. A more useful measure of model performance is provided by the "index of agreement" (Willmott, 1982; Willmott et al., 1985) which is defined as

$$d_\alpha = 1 - \frac{[\sum_{i=1}^{N} |P_i - O_i|^\alpha]}{[\sum_{i=1}^{N} (|P_i - \bar{O}| + |O_i - \bar{O}|)^\alpha]}$$

Table 2
The model predictor data for the MLP, regression tree and linear models. The standard meteorological input data for all the models, denoted MET in the table, consisted of [sin(wind direction), cos(wind direction), windspeed, amount of low cloud, visibility, dry bulb temperature, vapour pressure, base of lowest cloud]. d - Julian day of the year. h - hour of the day. The lower table indicates the amount of hourly training/validation/test data at each site for each model (missing data account for the variations in number of observations in each data set)

Model ID	Meteorological data	MLP architecture
01	MET	8:20:20:1
02	MET, $\sin(2\pi d/365)$, $\cos(2\pi d/365)$	10:20:20:1
03	MET, $\sin(2\pi d/365)$, $\cos(2\pi d/365)$, $\sin(2\pi h/24)$, $\cos(2\pi h/24)$	12:20:20:1

Site	Data
Bristol	15283/7802/7090
Edinburgh	15804/8375/8525
Eskdalemuir	16316/7322/8042
Leeds	16358/6315/7543
Southampton	13035/7653/7230

Table 3

Performance of all models when tested on independant test data. $\bar{O}$ – mean observed concentration (ppb). MBE - mean bias error (ppb). s_o and s_p - observed and predicted concentration standard deviation (ppb). MAE – mean absolute error (ppb). RMSE – root mean squared error (ppb). d_2 – index of agreement ($\alpha = 2$). a - linear best fit intercept (ppb). b - linear best fit gradient. Bootstrap estimates of standard error at the 95% confidence interval are listed in brackets. Models are described as either LM (linear model), RT (regression tree) or MLP (multilayer perceptron), with the model number indicating the predictor variables as described in Table 2

Statistic	LM01	LM02	LM03	RT01	RT02	RT03	MLP01	MLP02	MLP03
Bristol ($\bar{O}$ = 17.52(0.28))									
MBE	0.49(0.19)	0.35(0.20)	0.15(0.17)	0.65(0.18)	0.35(0.18)	0.14(0.18)	0.59(0.18)	0.14(0.17)	− 0.21(0.15)
s_o	11.67(0.24)	11.67(0.24)	11.67(0.24)	11.70(0.21)	11.70(0.21)	11.70(0.21)	11.68(0.25)	11.68(0.25)	11.68(0.24)
s_p	7.43(0.13)	7.42(0.12)	8.04(0.14)	8.08(0.19)	8.07(0.17)	8.56(0.16)	8.17(0.18)	8.83(0.22)	9.83(0.24)
MAE	6.59(0.12)	6.54(0.12)	6.13(0.11)	6.62(0.11)	6.63(0.12)	6.52(0.11)	6.24(0.11)	5.81(0.10)	5.09(0.10)
RMSE	8.30(0.16)	8.24(0.17)	7.70(0.14)	8.36(0.16)	8.40(0.17)	8.42(0.17)	7.84(0.14)	7.29(0.12)	6.60(0.13)
R^2	0.50(0.02)	0.51(0.02)	0.57(0.01)	0.49(0.02)	0.49(0.02)	0.48(0.02)	0.55(0.01)	0.61(0.02)	0.68(0.01)
d_2	0.79(0.01)	0.80(0.01)	0.83(0.01)	0.80(0.01)	0.80(0.01)	0.81(0.01)	0.83(0.01)	0.86(0.01)	0.90(0.01)
a	9.14(0.19)	9.22(0.20)	8.26(0.23)	8.20(0.21)	8.58(0.23)	8.30(0.23)	7.80(0.21)	7.02(0.24)	5.56(0.25)
b	0.45(0.01)	0.45(0.01)	0.52(0.01)	0.49(0.02)	0.48(0.01)	0.51(0.01)	0.52(0.01)	0.59(0.02)	0.69(0.01)
Edinburgh ($\bar{O}$ = 16.01(0.21))									
MBE	2.23(0.19)	2.21(0.18)	2.03(0.16)	2.23(0.19)	2.16(0.17)	2.08(0.16)	2.11(0.18)	2.33(0.16)	2.15(0.14)
s_o	9.71(0.15)	9.71(0.15)	9.71(0.15)	9.72(0.15)	9.72(0.15)	9.72(0.15)	9.71(0.15)	9.71(0.15)	9.71(0.15)
s_p	4.34(0.06)	4.85(0.06)	5.81(0.08)	5.09(0.10)	5.38(0.08)	6.43(0.08)	5.24(0.08)	6.16(0.09)	7.02(0.10)
MAE	6.86(0.12)	6.48(0.12)	6.14(0.11)	6.91(0.12)	6.45(0.12)	6.11(0.11)	6.48(0.11)	5.83(0.11)	5.11(0.10)
RMSE	8.86(0.18)	8.51(0.18)	7.98(0.17)	8.92(0.17)	8.46(0.17)	8.16(0.18)	8.31(0.16)	7.68(0.15)	6.80(0.13)
R^2	0.22(0.02)	0.28(0.01)	0.37(0.02)	0.21(0.01)	0.29(0.01)	0.35(0.01)	0.32(0.02)	0.43(0.02)	0.56(0.01)
d_2	0.58(0.01)	0.63(0.02)	0.71(0.01)	0.61(0.01)	0.67(0.01)	0.72(0.01)	0.67(0.01)	0.75(0.01)	0.83(0.01)
a	10.42(0.16)	9.53(0.18)	8.16(0.22)	9.89(0.18)	9.04(0.18)	7.69(0.22)	9.05(0.17)	7.01(0.19)	5.21(0.20)
b	0.21(0.01)	0.27(0.01)	0.36(0.01)	0.24(0.01)	0.30(0.01)	0.39(0.01)	0.30(0.01)	0.42(0.01)	0.54(0.01)
Eskdalemuir ($\bar{O}$ = 26.20(0.22))									
MBE	− 0.07(0.18)	− 0.25(0.17)	− 0.24(0.17)	− 0.13(0.18)	− 0.67(0.17)	− 0.66(0.17)	− 0.42(0.17)	− 0.44(0.15)	− 0.54(0.15)
s_o	10.41(0.18)	10.41(0.18)	10.41(0.18)	10.55(0.19)	10.55(0.19)	10.55(0.19)	10.39(0.19)	10.39(0.19)	10.39(0.1 8)
s_p	7.04(0.12)	7.49(0.12)	7.51(0.12)	8.75(0.17)	9.16(0.16)	9.16(0.16)	9.70(0.24)	9.75(0.25)	9.59(0.24)
MAE	6.41(0.11)	6.26(0.10)	6.26(0.10)	6.37(0.12)	6.21(0.11)	6.15(0.11)	5.84(0.10)	5.39(0.09)	5.38(0.09)
RMSE	8.20(0.15)	7.98(0.13)	7.98(0.14)	8.43(0.18)	8.11(0.16)	8.04(0.16)	7.63(0.15)	7.00(0.12)	6.93(0.12)
R^2	0.38(0.02)	0.42(0.01)	0.42(0.01)	0.40(0.02)	0.45(0.02)	0.46(0.02)	0.51(0.02)	0.58(0.02)	0.58(0.02)
d_2	0.75(0.01)	0.77(0.01)	0.77(0.01)	0.78(0.01)	0.81(0.01)	0.81(0.00)	0.84(0.01)	0.87(0.01)	0.87(0.01)
a	15.30(0.32)	14.25(0.35)	14.21(0.34)	12.44(0.43)	11.44(0.40)	11.30(0.40)	9.15(0.49)	7.91(0.49)	8.28(0.45)
b	0.42(0.01)	0.47(0.01)	0.47(0.01)	0.52(0.02)	0.58(0.02)	0.59(0.02)	0.67(0.02)	0.71(0.02)	0.70(0.02)

Table 3 (*continued*)

Statistic	LM01	LM02	LM03	RT01	RT02	RT03	MLP01	MLP02	MLP03
Leeds ($\bar{O}$ = 14.36(0.23))									
MBE	0.44(0.20)	0.58(0.19)	0.49(0.19)	0.18(0.18)	1.16(0.19)	1.01(0.18)	0.05(0.17)	0.40(0.17)	0.42(0.15)
s_o	10.76(0.26)	10.76(0.26)	10.76(0.26)	10.80(0.25)	10.80(0.25)	10.80(0.25)	10.78(0.26)	10.78(0.26)	10.78(0.26)
s_p	5.73(0.09)	5.88(0.09)	6.29(0.09)	6.84(0.18)	6.90(0.20)	7.30(0.20)	8.11(0.29)	7.76(0.21)	8.01(0.17)
MAE	6.69(0.13)	6.40(0.13)	6.22(0.12)	6.64(0.11)	6.64(0.11)	6.56(0.11)	5.86(0.11)	5.67(0.10)	5.01(0.10)
RMSE	8.68(0.21)	8.58(0.23)	8.23(0.21)	8.48(0.16)	8.65(0.18)	8.48(0.17)	7.49(0.15)	7.32(0.14)	6.55(0.15)
R^2	0.36(0.01)	0.37(0.02)	0.42(0.02)	0.38(0.02)	0.37(0.03)	0.40(0.02)	0.52(0.03)	0.54(0.02)	0.63(0.02)
d_2	0.69(0.01)	0.70(0.01)	0.74(0.01)	0.74(0.01)	0.74(0.01)	0.76(0.01)	0.83(0.01)	0.83(0.01)	0.87(0.01)
a	9.37(0.15)	9.00(0.17)	8.43(0.18)	8.36(0.21)	7.42(0.19)	7.08(0.21)	6.53(0.26)	6.36(0.20)	5.44(0.18)
b	0.32(0.01)	0.33(0.01)	0.38(0.01)	0.39(0.01)	0.39(0.02)	0.42(0.02)	0.54(0.02)	0.53(0.02)	0.59(0.01)
Southampton ($\bar{O}$ = 15.87(0.27))									
MBE	− 1.76(0.19)	− 1.83(0.18)	− 1.96(0.18)	− 1.71(0.20)	− 1.02(0.18)	− 1.24(0.19)	− 1.91(0.19)	− 1.33(0.17)	− 1.06(0.16)
s_o	11.06(0.19)	11.06(0.19)	11.06(0.19)	11.06(0.17)	11.06(0.17)	11.06(0.17)	11.07(0.20)	11.07(0.20)	11.07(0.19)
s_p	8.30(0.14)	9.00(0.14)	9.23(0.15)	9.21(0.25)	9.95(0.24)	9.87(0.23)	10.07(0.26)	10.28(0.24)	10.55(0.24)
MAE	6.72(0.11)	6.40(0.11)	6.30(0.11)	7.21(0.13)	6.49(0.12)	6.53(0.12)	6.52(0.12)	5.89(0.10)	5.22(0.10)
RMSE	8.27(0.13)	7.88(0.13)	7.80(0.13)	9.27(0.17)	8.55(0.18)	8.61(0.17)	8.37(0.15)	7.58(0.14)	6.83(0.14)
R^2	0.47(0.02)	0.53(0.02)	0.54(0.02)	0.37(0.02)	0.46(0.02)	0.45(0.01)	0.50(0.02)	0.57(0.01)	0.65(0.01)
d_2	0.80(0.01)	0.83(0.01)	0.84(0.01)	0.76(0.01)	0.82(0.00)	0.81(0.01)	0.83(0.01)	0.86(0.01)	0.89(0.01)
a	9.47(0.21)	8.32(0.23)	8.07(0.25)	9.32(0.26)	7.05(0.26)	7.40(0.25)	7.59(0.29)	6.04(0.25)	4.74(0.23)
b	0.51(0.02)	0.59(0.02)	0.62(0.02)	0.51(0.02)	0.61(0.02)	0.60(0.02)	0.64(0.02)	0.70(0.02)	0.77(0.02)

where P_i and O_i are the modelled and observed concentrations. The index of agreement can be calculated for both $\alpha = 1$ and 2. Both values of d_α are normalised and indicate the extent that predicted deviations differ from the observed deviations about the mean observed value, indicating the degree to which model predictions are error free. The d_2 statistic is more sensitive than the d_1 statistic since it is based on simple differences rather than squared differences. Since d_α is a bounded and relative measure it can be used to make comparisons between models. Also reported are the linear best fit coefficients a and b given by the linear model $P_i = a + bO_i$.

Figs. 1–4 display the relative performance of the linear, regression tree and MLP models at each site in terms of $(s_o - s_p)$, RMSE, R^2 and d_2.

7. Discussion

The MBE values listed in Table 3 exhibit very little variability at a site, as both the modelling techniques and predictor variables change. At Bristol and Leeds the MBE is generally small and positive indicating a tendency to underpredict concentrations. At Eskdalemuir the MBE is generally small and negative indicating slight overprediction by the models. In contrast at Edinburgh the MBE is much larger (around 2 ppb) and indicates a general underprediction whilst at Southampton the MBE indicates a similar level of overprediction. The observed mean concentrations are of similar value at all the urban sites hence this is unlikely to explain the differences in MBE. At Edinburgh the most probable explanation is that the meteorological data were collected at the airport 10 km to the west of the city centre where the ozone data were collected. The representativeness of the meteorological data is questionable and likely leads to the poorer model performance. At Southampton the meteorological and ozone data were collected, spatially, much closer together. However this model was developed and tested over a different time period. The range of concentrations observed during the training and validation data at Southampton were between 1 and 93 ppb with a median of 15 ppb. During

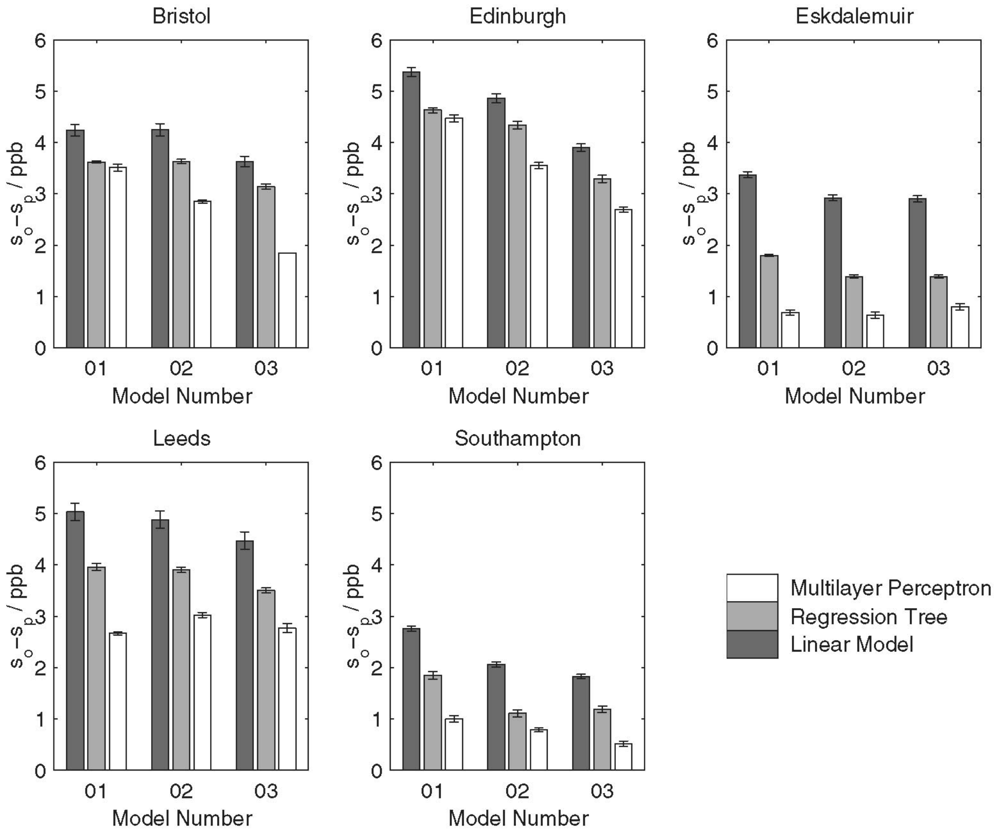

Fig. 1. Relative performance (in terms of $(s_o - s_p)$) of linear, regression tree and multilayer perceptron models at each site. Model numbers indicate the predictor variables used, as described in Table 2. Error bars indicate standard error in s_p at 95% confidence level.

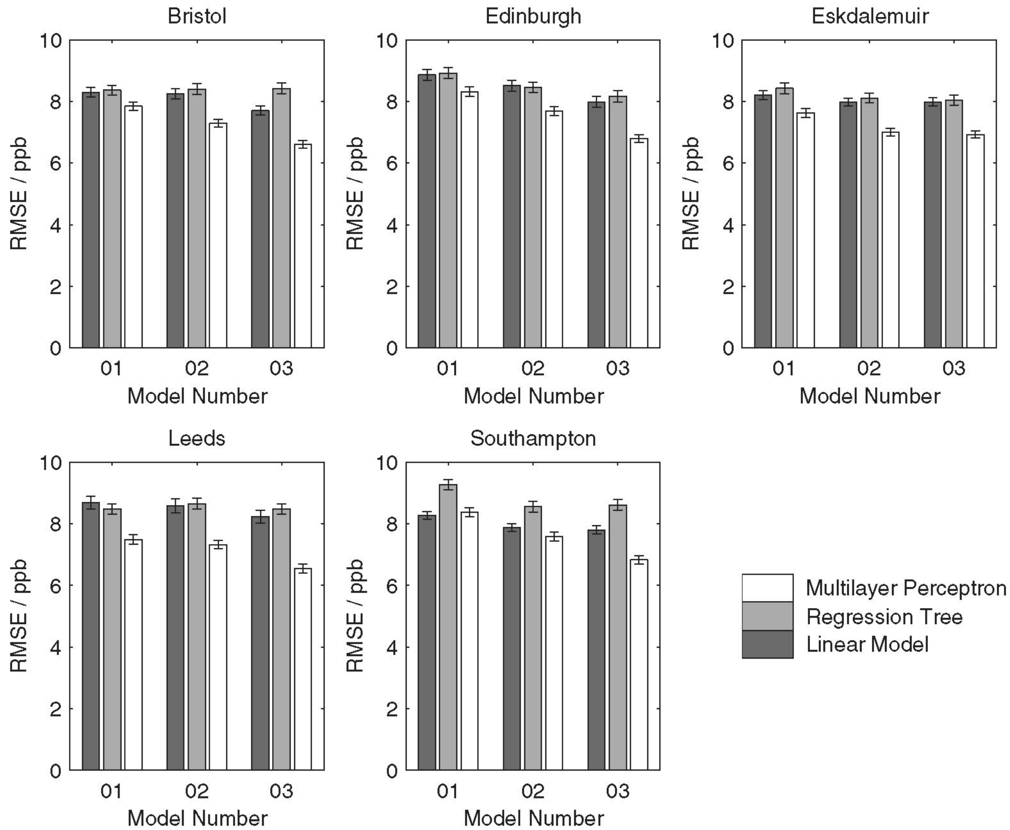

Fig. 2. Relative performance (in terms of RMSE) of linear, regression tree and multilayer perceptron models at each site. Model numbers indicate the predictor variables used, as described in Table 2. Error bars indicate standard error at 95% confidence level.

the test data period the concentrations ranged from 1 to only 77 ppb with a median of 14 ppb. This could be causing the overpredictions observed at this site.

Fig. 1 shows that there is a clear ranking of the modelling methodologies, in terms of $(s_o - s_p)$, with the linear model producing concentrations with much less variability than the observed data. The regression tree and MLP modelled concentrations display variability which increasingly matches the observed variability. In all cases the observed variance is greater than the modelled variance, a feature typical of statistical models which attempt to model the general behaviour evident in the data. Highly constrained models, such as the linear models, have reduced flexibility restricting the ability of the model to capture the original variability in the data. The effect of adding the time of day and time of year inputs to the MLP models, for Eskdalemuir and Leeds, does not appear to improve the models, in terms of $(s_o - s_p)$. Eskdalemuir will be discussed later, however in terms of the other statistical measures, the inclusion of time of day and time of year in the MLP model for Leeds results in an improved model. This demonstrates the importance of considering more than one statistical measure when assessing model performance.

In terms of RMSE or R^2, as shown in Figs. 2 and 3, the linear and regression tree models exhibit similar levels of performance, with the MLP models performing best at all sites and with all combinations of input variables. MAE, listed in Table 3 and not shown graphically, parallels the behaviour of RMSE. The similar performance of the regression tree and linear models was unexpected, however this is most likely due to the piecewise representation of a function in a regression tree. Within the terminal nodes of the regression tree the modelled concentrations are assumed constant even though observed concentrations may exhibit considerable variability. This behaviour will introduce additional model errors. At Southampton the regression trees are seen to perform significantly worse than the linear models suggesting that these trees were either "over – pruned" and therefore incapable of accurately representing the relationship between meteorology and ozone, or "under-pruned" and overfitting the relationship resulting in poor generalisation capabilities. However, tests of both these hypotheses

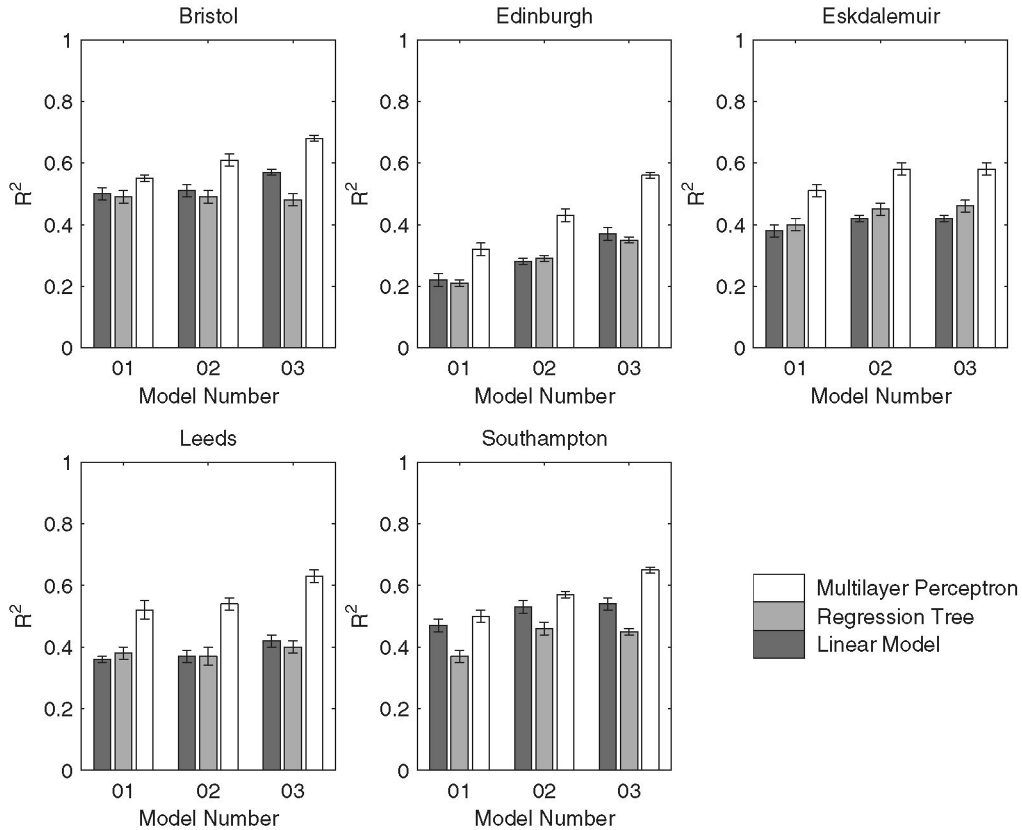

Fig. 3. Relative performance (in terms of R^2) of linear, regression tree and multilayer perceptron models at each site. Model numbers indicate the predictor variables used, as described in Table 2. Error bars indicate standard error at 95% confidence level.

indicate that the chosen tree has the best generalisation performance, and suggest that the relationship between ozone and meteorology cannot be accurately represented in a tree-based structure using these predictor variables at Southampton.

The best fit line through the observed and modelled concentrations provides another method to test model performance. More accurate models will have intercepts (a) tending to 0 and gradients (b) tending to 1. Table 3 shows that the MLP models have intercepts closest to the origin and gradients nearest to unity. The regression tree and linear model best fits become increasingly and significantly worse. Fig. 5 shows scatterplots of modelled and observed concentrations during the test period, obtained from the various models with meteorological, time of year and time of day inputs (model 03). It can be seen that whilst the MLP model resolves the higher concentrations more accurately, underpredictions do occur.

The index of agreement, shown in Fig. 4, again indicates that the MLP model is performing best. At all sites except Southampton, and to a lesser degree at Bristol, the worst performance is obtained from the linear models. Generally, regardless of statistical measures and for all modelling techniques, model 03 performs better than model 02 which in turn is better than model 01. This indicates that the time of year and time of day are useful hourly ozone predictor variables.

In contrast to the results obtained by Comrie (1997), where it was noted that differences in performance of the linear and MLP models of daily maximum ozone often lay within the error bounds, these results clearly demonstrate significant improvements when using MLP models. As an example Table 4 shows the difference in the performance of the MLP and the linear models, developed here, in terms of R^2. For the models with only meteorological inputs at Leeds, Eskdalemuir, and Edinburgh the increase in the explained variance is substantial, amounting to between 10–15%. Only at Bristol and Southampton is the increase relatively small. As the seasonal and time of day inputs are added to the models the difference in performance increases further to between 11 and 21%. This explainable variance cannot be captured by the linear models since the relationship involves nonlinear interactions between the predictor variables and

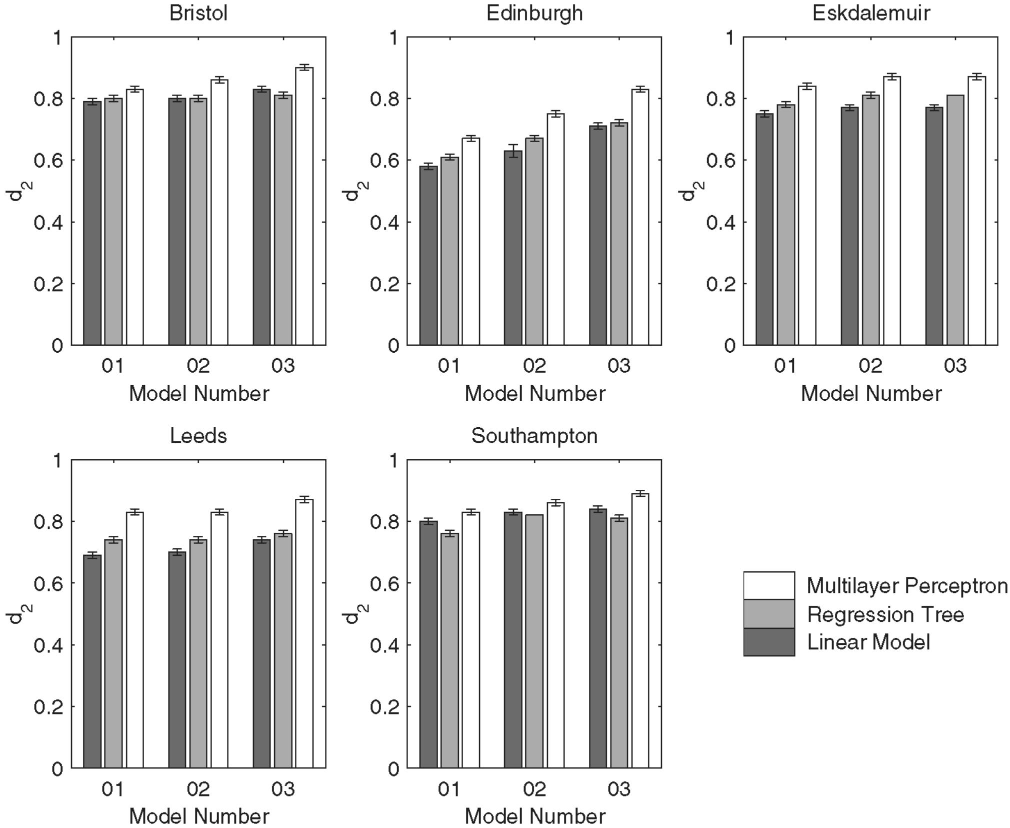

Fig. 4. Relative performance (in terms of d_2) of linear, regression tree and multilayer perceptron models at each site. Model numbers indicate the predictor variables used, as described in Table 2. Error bars indicate standard error at 95% confidence level.

nonlinear relationships between predictors and predictand. This clearly demonstrates the need to use an appropriate model when simulating hourly surface ozone concentrations; inappropriate models could result in the role of meteorology in determining surface ozone concentrations being significantly understated.

The site-specific performance of the linear ozone models is symptomatic of the varying complexity of the relationship between ozone and meteorology between sites. Table 4 indicates that a linear model would make a better approximation, using these predictor variables, at Bristol or Southampton than at Edinburgh, Eskdalemuir or Leeds. When comparing models the index of agreement (d_2) is a useful statistic since it is bounded and unrelated to the observed mean concentration. Comparing the performance of the MLP and linear models in terms of d_2 indicates that there is much less variability in the performance of the MLP models ($0.83 \leqslant d_2 \leqslant 0.90$) than the linear models ($0.71 \leqslant d_2 \leqslant 0.84$). The variability in performance at each site is a function of the representativeness and suitability of the meteorological data ("data issues") and the choice of model. The variation in performance of the MLP models is dominantly related to "data issues" and is not due to the choice of model. With inappropriate models, for example inflexible linear models, such an attribution of the variation in performance of the model to data is confounded by problems with the underlying model itself.

It has already been noted that the representativeness of the meteorological data at Edinburgh may have been contributing to the relatively poor performance of the models developed at this site. The addition of seasonal, and especially time of day inputs, to the MLP models decreases the difference in performance between the Edinburgh model and the models for the other sites. The time of day input was added to enable the models to compensate for the temporal variation in NO_x emissions which act to chemically destroy ozone concentrations. However, at Edinburgh it is also possible that the time of day inputs are allowing the model to compensate for the unrepresentative meteorological data by adjusting the relationship between the meteorological data and ozone concentrations throughout the day. The increase in R^2 at Edinburgh, with the addition of time of day inputs, is 0.13 compared with between 0.00 (Eskdalemuir) and 0.09

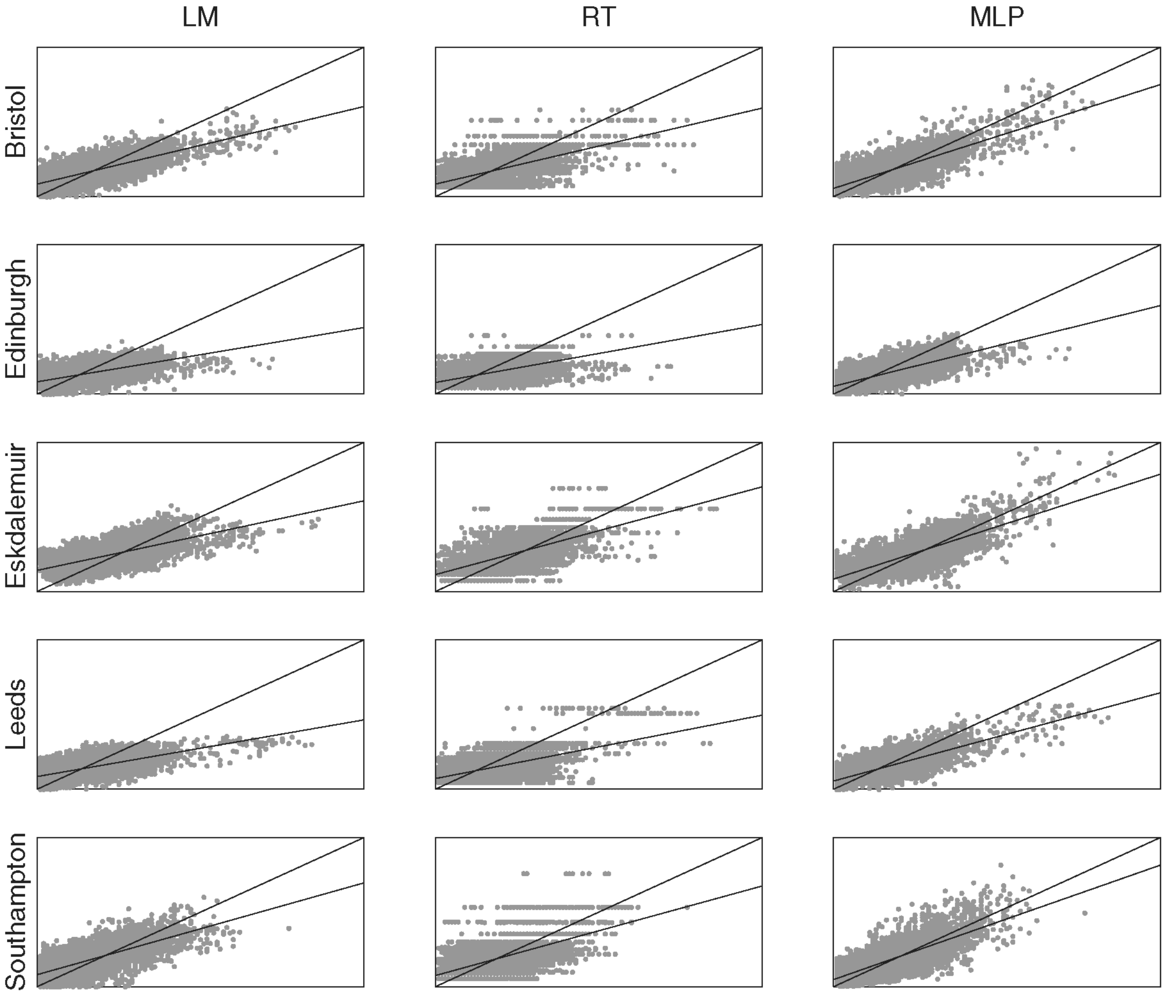

Fig. 5. Scatterplot of observed (x-axis) and modelled (y-axis) concentrations for linear (LM), regression tree (RT) and multilayer perceptron (MLP) models 03. Both axes range from 0–100 ppb. Best fit and least squares lines shown.

Table 4
The difference in performance of the MLP and the linear models in terms of R^2. Difference calculated as (MLP Model - Linear Model). Models described in Table 2. Bootstrap estimates of a statistically significant increase at the 95% confidence interval listed in brackets

	01	02	03
Bristol	0.05(0.03)	0.10(0.04)	0.11(0.02)
Edinburgh	0.10(0.04)	0.15(0.03)	0.19(0.03)
Eskdalemuir	0.13(0.04)	0.16(0.03)	0.16(0.03)
Leeds	0.16(0.04)	0.17(0.04)	0.21(0.04)
Southampton	0.03(0.04)	0.04(0.03)	0.11(0.03)

(Leeds) at the other sites. Interestingly at Eskdalemuir, where NO_x concentrations would be expected to be low and to exhibit little diurnal variation due to the remote location of the site, there is no increase in model performance with the addition of the time of day inputs.

The addition of the seasonality inputs causes a greater improvement in the performance of the MLP models than in the linear models. The nonlinear interactions that can occur between the input variables in the MLP models are responsible. The relationship between time of day, time of year and surface ozone concentrations cannot be represented by a linear function. The MLP models are not constrained in this way and make more use of these inputs, providing the model with greater freedom to capture the underlying surface ozone behaviour.

7.1. Regression tree model interpretation

Although the regression tree models do not perform as well as the MLP models and only marginally outperform

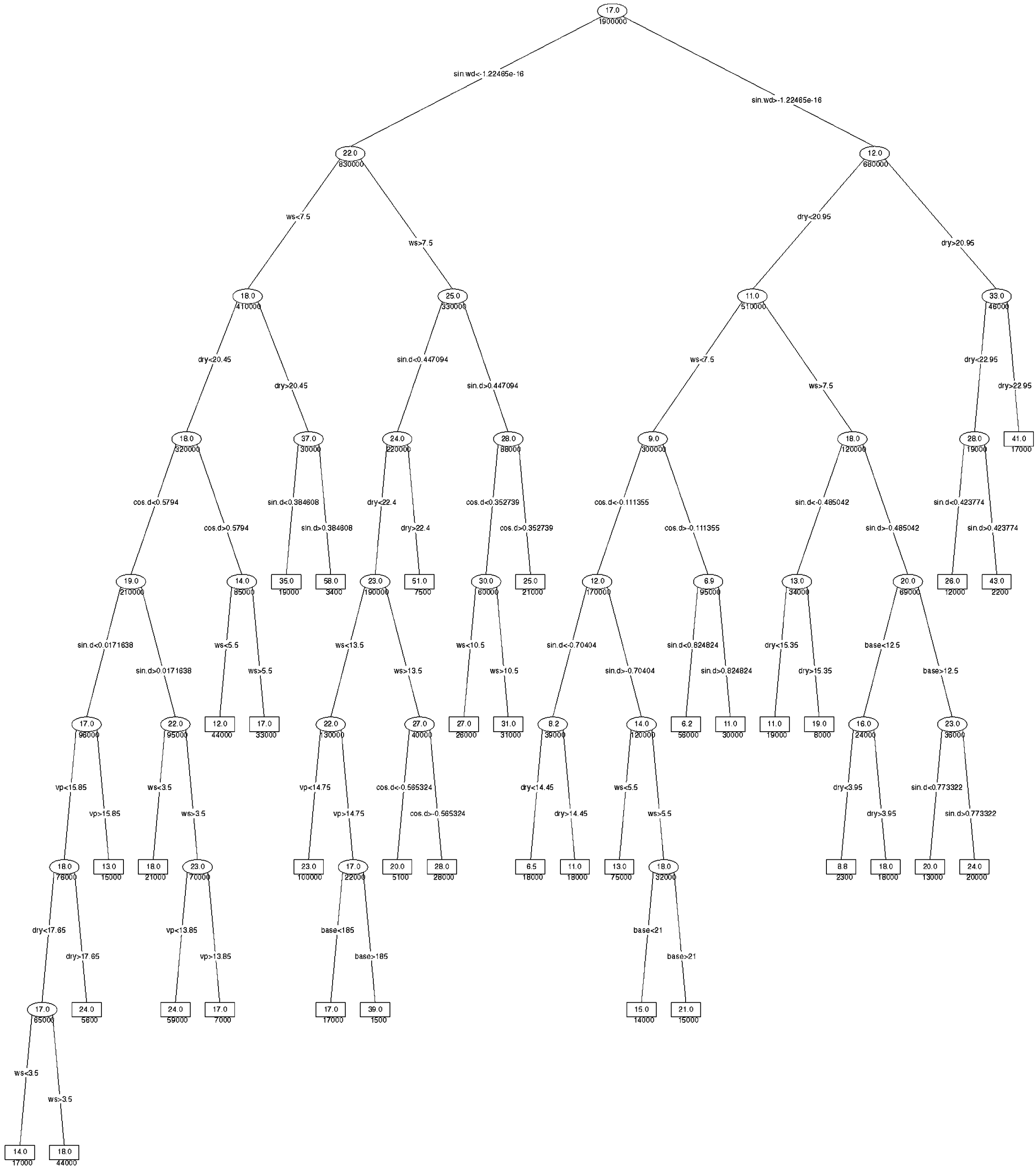

Fig. 6. Bristol regression tree model (02). The edges connecting uniformly spaced nodes are labeled with the left and right splitting rules. Interior nodes are represented by ellipses, whilst terminal nodes are represented by rectangles. All nodes are labelled with their values and the within-node variance is printed below each node. Abbreviations are as follows: dry - temperature (°C), vp – vapour pressure (mb), ws – windspeed (knots), wd – wind direction, base – base of lowest cloud (decameters), vis – visibility (m), sin.d – $\sin(2\pi d/365)$ where d is Julian Day, cos.d - $\cos(2\pi d/365)$ where d is Julian Day.

the linear models, they have the advantage of being readily interpretable. Conversely, coefficients in regression equations require careful interpretation (Sowizral and Zurbenko, 1998) and at present no reliable methods exist to interpret MLP models. As an example to demonstrate the transparency of the regression tree models, the tree for Bristol using meteorological and time of year predictor variables is presented in Fig. 6 and described here. Similar analyses of the other regression trees have been made (Gardner, 1999) however space limitations restrict their inclusion here.

The first split in the Bristol tree (Fig. 6) uses wind direction to divide the data into an easterly and westerly regime. For the westerly regime (the left side of the tree) the range in ozone concentrations observed in the terminal nodes is between 12 and 58 ppb whilst for the easterly regime the concentrations observed in the terminal nodes range from 6 to 43 ppb. Intuitively it might be expected that highest concentrations would occur when the wind is from an easterly direction with air having travelled over the UK and possibly the continent entraining both ozone and its precursors. Analysis of the distribution of ozone concentrations observed within both the easterly and westerly regimes indicate that the maximum concentrations occur when the temperatures are high during spring and early summer in both regimes. However with easterly winds the maximum concentration (73 ppb) is lower than that observed in the westerly regime (89 ppb). This may indicate that by the time air arrives in Bristol, from an easterly direction, chemical destruction of ozone due to reactions with NO emissions may have occured. The location of the monitoring site, much closer to the southwest edge of the city than the eastern boundary, is consistent with such an explanation. In contrast, during anticyclonic conditions which result in the local wind direction at Bristol being from a westerly direction, for example when an anticyclonic circulation system is centred to the south and east of the UK, such interactions with NO emissions would be limited leading to higher ozone concentrations.

Within the westerly regime windspeed has been selected to further split the data. When windspeeds are low (less than 7.5 knots) the maximum concentration (58 ppb) occurs when temperatures are greater than 20.5°C during springtime. The identification of springtime as one of the splitting rules is interesting since it is well known that there is a springtime maxima in surface ozone concentrations (Derwent et al., 1998). Springtime is represented by splitting rules involving $\sin(d)$ being greater than zero. The high temperatures also indicate the possible importance of photochemistry. With slightly higher windspeeds the concentrations within this westerly regime are lower (51 ppb) suggesting that higher windspeeds are not generally associated with strong photochemistry (although they may be important for vertical transport).

In the easterly regime the tree uses temperature to split the data further. When temperatures are high (greater than 20.9°C) the concentrations are in the range 26–43 ppb, whereas with lower temperatures the concentrations are in the range 6–24 ppb. These rules again indicate the detection of a photochemical signal for which temperature is a good surrogate.

8. Conclusion

In contrast with Comrie (1997), where only slight increases in performance between linear and MLP models of daily maximum ozone were observed, these results, using more detailed input at the hourly timescale, demonstrate that significant increases in performance are possible when using MLP models. The complexity of the relationship between the meteorological predictor variables and hourly ozone concentrations was seen to vary between sites.

The regression tree models were seen to be more readily interpretable at the expense of reduced model performance. Analysis of the regression trees demonstrated that the models had identified mechanisms that made physical sense.

The MLP models performed best since they were unconstrained and allowed arbitrary interactions and nonlinear relationships between predictor variables. Unfortunately the MLP models cannot easily be interpreted, although various sensitivity tests and model comparisons may provide insight into model behaviour. However for prediction purposes it could be argued that a good "black box" model is better than a poor, yet well understood, physically based model. The improved statistical models of surface ozone have enabled long term trends in surface ozone concentrations (due to changes in precursor emissions) to be addressed both in the US (Gardner and Dorling, 1999b) and UK (Gardner and Dorling, 1999c).

Acknowledgements

We are grateful to the BADC and DETR/NETCEN for making the meteorological and ozone data freely available and also to the School of Environmental Sciences, University of East Anglia, for supporting this work.

References

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Bloomfield, P., Royle, J.A., Steinberg, L.J., Yang, Q., 1996. Accounting for meteorological effects in measuring urban ozone levels and trends. Atmospheric Environment 30 (17), 3067–3077.

Breiman, L., Freidman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth and Brooks-Cole. Belmont, CA.

Burrows, W.R., Benjamin, M., Beauchamp, S., Lord, E., McCollor, D., Thompson, B., 1995. CART decision – tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic Regions of Canada. Journal of Applied Meteorology 34, 1848–1862.

Comrie, A.C., 1997. Comparing neural networks and regression models for ozone forecasting. Journal of the Air and Waste Management Association 47, 653–663.

Davies, T.D., Kelly, P.M., Low, P.S., Pierce, C.E., 1992. Surface ozone concentrations in Europe: links with the regional-scale atmospheric circulation. Journal of Geophysical Research 97 (D9), 9819–9832.

Derwent, D.G., Davies, T.J., 1994. Modelling the impact of NO_x or hydrocarbon control on photochemical ozone in Europe. Atmospheric Environment 28 (12), 2039–2052.

Derwent, R.G., Simmonds, P.G., Seuring, S., Dimmer, C., 1998. Observation and interpretation of the seasonal cycles in the surface concentrations of ozone and carbon monoxide at Mace Head, Ireland from 1990 to 1994. Atmospheric Environment 32 (2), 145–157.

Entwistle, J., Weston, K., Singles, R., Burgess, R., 1997. The magnitude and extent of elevated ozone concentrations around the coasts of the British Isles. Atmospheric Environment 31 (13), 1925–1932.

Feister, U., Balzer, K., 1991. Surface ozone and meteorological predictors on a subregional scale. Atmospheric Environment 25A (9), 1781–1790.

Fox, D.G., 1981. Judging air quality model performance: a summary of the AMS Workshop on Dispersion Model Performance. Bulletin of the American Meteorological Society 62, 599–609.

Gardner, M.W., 1999. The advantages of artificial neural network and regression tree based air quality models, Ph.D. Thesis, School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, NR4 7TJ, UK.

Gardner, M.W., Dorling, S.R., 1996. Neural network modelling of the influence of local meteorology on surface ozone concentrations. In: Abrahart, R.J. (Ed.), Proceedings first International Conference on GeoComputation, Vol. 1, pp. 359–370.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences. Atmospheric Environment 32 (14–15), 2627–2636.

Gardner, M.W., Dorling, S.R., 1999a. Neural network modelling of hourly NO_x and NO_2 concentrations in urban air in London. Atmospheric Environment 33 (5), 709–719.

Gardner, M.W., Dorling, S.R., 1999b. Artificial neural network derived trends in surface ozone concentrations. Journal of the Air and Waste Management Association. Submitted for publication.

Gardner, M.W., Dorling, S.R., 1999c. Meteorologically adjusted trends in UK daily surface ozone concentrations. Atmospheric Environment, accepted for publication.

Hornik, K., Stinchcoombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366.

Hubbard, M.C., Cobourn, W.G., 1998. Development of a regression model to forecast ground-level ozone concentration in Louisville, KY. Atmospheric Environment 32 (14–15), 2637–2647.

Kunz, H., Speth, P., 1997. Variability of near-ground ozone concentrations during cold front passages – a possible effect of tropopause folding events. Journal of Atmospheric Chemistry 28, 77–95.

Lim, T.S., Loh, W.Y., Shih, Y.S., 1997. An empirical comparison of decision trees and other classification methods. In: Technical Report 979, Department of Statistics. University of Wisconsin, Madison.

Moller, M.F., 1991. A scaled conjugate gradient algorithm for fast supervised learning. IEEE Transactions on Systems, Man and Cybernetics 21, 272–280.

Moller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 6, 525–533.

O'Hare, G.P., Wilby, R., 1995. A review of ozone pollution in the united kingdom and ireland with an analysis using lamb weather types. The Geographical Journal 161 (1), 1–20.

PORG, 1997. Ozone in the United Kingdom, Photochemical Oxidants Review Group.

Sarle, W.S., 1995. Stopped training and other remedies for overfitting. In: Proceedings of the 27th Symposium on the Interface. Available via ftp://ftp.sas.com/pub/neural/inter95.ps.Z.

Sowizral, M., Zurbenko, I.G., 1998. The destructive effect of noise in linear regression. Journal of Statistical Computation and Simulation, submitted for publication.

Venables, W.M., Ripley, B.D., 1994. Modern applied statistics with S-plus. In: Statistics and Computing. Springer, Berlin.

Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society 63 (11), 1309–1313.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. Journal of Geophysical Research 90 (C5), 8995–9005.

Yi, J., Prybutok, R., 1996. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environmental Pollution 92 (3), 349–357.