# Maximum Likelihood Cost Functions for Neural Network Models of Air Quality Data.

Stephen R. Dorling [a] Robert J. Foxall [a] Danilo P. Mandic [b]
Gavin C. Cawley [c,*]

[a] *School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom.*

[b] *Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, London SW7 2BT, United Kingdom.*

[c] *School of Information Systems, University of East Anglia, Norwich NR4 7TJ, United Kingdom.*

**Abstract**

The prediction of episodes of poor air quality using artificial neural networks is investigated, concentrating on selection of the most appropriate cost function used in training. Different cost functions correspond to different distributional assumptions regarding the data, the appropriate choice depends on whether a forecast of absolute pollutant concentration or prediction of exceedence events is of principle importance. The cost functions investigated correspond to logistic regression, homoscedastic Gaussian (i.e. conventional sum-of-squares) regression and heteroscedastic Gaussian regression. Both linear and non-linear neural network architectures are evaluated. While the results presented relate to a dataset describing the daily time-series of the concentration of surface level ozone ($O_3$) in urban Berlin, the methods applied are quite general and applicable to a wide range of pollutants and locations. The heteroscedastic Gaussian regression model outperforms the other non-linear methods investigated, however there is little improvement resulting from the use of non-linear rather than linear models. Of greater significance is the flexibility afforded by the non-linear heteroscedastic Gaussian regression model for a range of potential end-users, who may all have different answers to the question: "What is more important, correctly predicting exceedences or avoiding false alarms?".

* Corresponding author, tel: +44 (0)1603 593258, fax: +44 (0)1603 592245, email: gcc@sys.uea.ac.uk

[1] This work was supported by the European Commission, grant number IST-99-11764, as part of its Framework V IST programme and by the Royal Society, grant number RSRG-22270.

## 1 Introduction

In recent years, neural network models have been widely developed and applied to atmospheric science problems in general (Gardner and Dorling, 1998) and air quality problems in particular (Comrie, 1997; Gardner and Dorling, 1999a; Gardner and Dorling, 1999b). Advantages of neural network based approaches include the ability to simulate non-linear behaviour and to avoid the necessity of making unnecessary and potentially incorrect assumptions regarding interactions between model input variables. The APPETISE project (Grieg, 2000), funded under the IST programme of the European Commission's Framework V programme, was conceived to comprehensively test the ability of neural network models applied to air quality problems and this paper reports on an informative subset of the findings of this project.

Episodes of elevated concentrations of air pollutants, which are thought to be potentially damaging to health, are addressed in the European Air Quality Framework Directive and subsequent daughter directives. These directives identify threshold concentrations and pollutant-specific periods over which these concentrations occur which are thought to be associated with these deleterious health effects. This paper specifically addresses how neural network models perform in simulating pollutant concentrations, and importantly prediction of exceedences of statutory threshold concentrations during these episodes. A vital issue that has not been addressed in previous studies is that for most end-users the costs associated with false-positive errors (i.e. false-alarms) and false-negative errors (i.e. failing to predict the occurrence of an observed exceedence) are unlikely to be equal, and furthermore are likely to differ widely for different groups of end-users. The aim of this study is to develop neural network models of air pollutant datasets that give accurate predictions, but also give flexibility in accommodating unequal misclassification costs.

The meteorological conditions which can help drive the incidence of episodes of poor air quality include low wind speeds and temperature inversions (poor ventilation), solar radiation intensity (production of photochemical pollutants) and external air temperature (related to pollutant emission rates from industrial plant, heating systems and vehicular activity). Some of the meteorological and pollutant emission variability is also, of course, captured to some extent by simple temporal predictor variables such as "month of the year", "day of the week" and "time of day". The neural network models described here therefore combine input parameters describing the recent history of pollutant concentrations, observed meteorological data and temporal information. In this paper we inter-compare cost functions representing different distributional assumptions regarding the target data.

Neural networks provide a flexible, non-linear model relating these predictor variables to pollutant concentrations. However, it has often been observed (Bishop, 1995) that simple maximum likelihood estimates for the parameters of a complex statistical model, such as a multi-layer perceptron neural network, often lead to severe over-fitting of the training data. For practical applications it is therefore vital to limit the complexity of the model to suit the complexity of the learning problem defined by the data. The two most commonly encountered approaches are known as formal regularisation and structural stabilisation. The latter approach, in the case of the multi-layer perceptron, seeks to determine the optimal number of hidden neurons and synapses, either by pruning a large network and retraining or starting with a simple model and adding resources while generalisation continues to improve (Hassibi and Stork, 1993). Formal regularisation, on the other hand, incorporates a regularisation term into the cost function that seeks to penalise overly complex models. In this study we adopt a Bayesian regularisation scheme due to (Williams, 1995) that provides both formal regularisation and structural stabilisation.

## 2 Unequal Misclassification Costs

It is rare in practical applications of statistical pattern recognition algorithms, including artificial neural networks, for the cost of false-positive and false-negative misclassification errors to be the same. Consider, for example, a medical screening test, where a patient is classified as suffering from a particular disease (a positive diagnosis) or being free of that disease (a negative diagnosis) based on some set of physiological measurements. In this case, a false-positive error, while understandably the source of some considerable anxiety for the patient, is relatively benign as the patient will be referred back to the doctor for more sophisticated tests likely to reveal the error. On the other hand, a false-negative error is potentially far more serious as the disease may become more advanced before the error is noticed, putting the patient at greater risk and also increasing the eventual expense of treatment. Ideally the costs of false-positive and false-negative misclassification errors should be factored into the design of the classifier.

In the case of a binary (two class) pattern recognition system, that gives as an output the probability, $p(\vec{x})$, that the input pattern, $\vec{x}$, belongs to class $\mathcal{A}$ rather than class $\mathcal{B}$, it is straight-forward to accommodate unequal misclassification costs. The input pattern is normally assigned to class $\mathcal{A}$ if $p(\vec{x}) \geq \tau$ and to $\mathcal{B}$ if $p(\vec{x}) < \tau$; if the misclassification costs are equal then $\tau = 0.5$. It is easy to show that the use of unequal misclassification costs corresponds to a threshold given by $\tau = c_{fp}/(c_{fn} + c_{fp})$, where $c_{fn}$ and $c_{fp}$ are the costs associated with false-negative and false-positive errors respectively.

In the case of statistical prediction of episodes of poor air quality, the situation is complicated by the fact that different end-users are likely to have different, possibly contradictory, views on the correct set of misclassification costs. For instance a hospital manager may wish to balance the cost of additional staffing during episodes of poor air quality to cope with an influx of patients expe-

riencing breathing difficulties against the cost of more expensive treatment of seriously ill patients if immediate care is not available. Alternatively a city council may act to close the city centre to traffic on a day where an exceedence of a statutory threshold is predicted, perhaps in order to avoid being penalised by a higher level of government, but can only achieve this at the expense of a reduction in city centre trade. While, in these hypothetical scenarios, the costs have been formulated in financial terms, it is not always possible to derive the true misclassification costs via a rigorous numerical procedure and one must instead rely on expert opinion. The important concept presented here is that if an end-user is to act on a prediction made by an air pollution model, it is vital to ensure that the appropriate misclassification costs are applied, firstly to minimise the expected loss, but also to ensure that any action or inaction can be properly justified.

## 3 Neural Models of Air Pollution Time-Series

In this paper, we inter-compare five approaches to predicting episodes of poor air quality using artificial neural networks and generalised linear models. Three different cost functions are evaluated corresponding to logistic regression, homoscedastic Gaussian regression and heteroscedastic Gaussian regression. Homoscedasticity implies that the variance of the target distribution is independent of the explanatory variables, a common assumption that is not always justified. For instance, one could argue that the variability in ozone concentrations is higher in Summer than during Winter months. One possible justification for this assumption would be the observation that the absolute amount of direct sunlight reaching the Earth's surface, a factor directly influencing ozone concentrations, is not only higher in Summer but also more variable. The Winter months, at least in the United Kindom, are predominantly overcast, whereas during Summer one observes an alternating pattern of bright and cloudy spells. In this situation, it would be more appropriate to employ

a heteroscedastic distribution, where both the mean and variance of the target distribution are modelled as a function of the explanatory variables, in this case time of year. In this study we have used Generalised Linear Models (GLMs) (Nelder and Wedderburn, 1972) and Multi-Layer Perceptron (MLP) neural networks, although the methods used are equally applicable to Radial Basis Function neural networks (Bishop, 1995). In the remainder of this paper, we assume that the reader is familiar with these models and highlight only the specific features required to accommodate different cost functions used in training.

### *3.1 Logistic Regression*

Logistic regression provides perhaps the most straight-forward approach to predicting episodes of poor air quality. Assuming the target patterns, $t_i$, represent an independent and identically distributed (i.i.d.) sample drawn from a Bernoulli distribution ($t_i = 1$ indicates an exceedence, $t_i = 0$ indicates the absence of an exceedence), conditioned on the corresponding input vectors, $\vec{x}_i$, the *likelihood* of the training data $\mathcal{D}$, given a vector of model parameters $\vec{w}$, is given by

$$p(\mathcal{D} \mid \vec{w}) = \prod_{i=1}^{n} (y_i)^{t_i} (1 - y_i)^{1-t_i}, \tag{1}$$

where $y_i$ is the output of the model for the $i^{th}$ observation. Taking the negative logarithm gives rise to the familiar *cross-entropy* cost function (Bishop, 1995),

$$E_{\mathcal{D}} = -\sum_{i=1}^{n} \{t_i \log y_i + (1 - t_i) \log(1 - y_i)\}. \tag{2}$$

The output of a linear model, $\eta_i = \vec{w}\vec{x}_i$ is related to the probability of an exceedence via the *logit* link function:

$$\eta_i = \text{logit}(y_i) \quad \Leftrightarrow \quad y_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}.$$

Maximum likelihood estimates for these parameter values are found using an iteratively weighted least squares algorithm (Nelder and Wedderburn, 1972).

For non-linear logistic regression models implemented using multi-layer perceptron networks, the output layer activation function is chosen to be the sigmoidal logistic function, $g(a) = 1/(1 + e^{-a})$, restricting the output of the model to lie in the range $(0,\ 1)$. Under these conditions the output of the model provides a maximum likelihood estimate of the Bayesian *a-posteriori* probability of an exceedence (Hopfield, 1987). Note the logistic regression models do not provide an estimate of the concentration of a given pollutant, but only the probability that this concentration exceeds a predetermined threshold.

### *3.2 Homoscedastic Gaussian Regression*

The use of a sum-of-squares error (SSE) cost function (3) corresponds to maximum likelihood estimation of the conditional mean of the target distribution, assuming a homoscedastic Gaussian noise process. The assumption of constant variance simplifies the cost-function, giving

$$E_D = \sum_{i=1}^{N}(y_i - t_i)^2 \tag{3}$$

For linear sum-of-squares regression, where the output of the model is given by $y_i = \vec{w}\vec{x}_i$, the optimal parameter values are found directly by solving the well-known *Normal equations* (Weisberg, 1985). For non-linear models implemented using neural networks, there is a single output unit with a linear activation function $g(a) = a$. The output of the model can be regarded as specifying a distribution for the predicted pollutant concentration, known as the predictive distribution, in this case a Gaussian distribution centered on the conditional mean given by the output of the model, $y_i$, with variance given by the usual maximum likelihood estimate,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2.$$

The probability of an exceedence is then given by the integral of the predictive distribution above the fixed exceedence threshold. This is illustrated by Figure

1, where the probability of an exceedence is given by the area of the shaded region. An episode of poor air quality is then predicted should this probability exceed a fixed threshold, as described in section 2.

### *3.3 Heteroscedastic Regression*

A heteroscedastic Gaussian regression model relaxes the strong assumption that the noise process is of constant variance (i.e. independent of the input pattern $\vec{x}_i$), and so attempts to estimate both the conditional mean and conditional variance of the target distribution (Nix and Weigand, 1995; Williams, 1996), providing a more flexible and potentially realistic model. For a Gaussian noise process, the network then has two output units, $y_i^\mu$ estimating the conditional mean of the target distribution and $y_i^\sigma$ estimating the conditional standard deviation. The negative logarithm of the likelihood of the data is then given by

$$-\log p(\mathcal{D} \mid \vec{w}) = \frac{1}{2}\sum_{i=1}^{N}\left\{\frac{[t_i - y_i^\mu]^2}{(y_i^\sigma)^2} + \log(y_i^\sigma)^2 + \log 2\pi\right\}, \tag{4}$$

which can be used to form a maximum likelihood cost function as before. Again, the probability of an exceedence is given by the integral of the predictive distribution above the exceedence threshold, except that the variance is now also estimated by the model.

### *3.4 Linear and Non-linear Models*

A generalised linear model can be considered as a special case of the multi-layer perceptron neural networks, discarding the layer of hidden units. A failure of the non-linear model to outperform its linear counterpart could be due to the data having a linear or nearly linear structure, or to the non-linear networks being poorly trained. Linear models are therefore used to test the assumption that a non-linear model is required for accurate prediction of pollutant levels.

## 4 Results

The dataset used to illustrate the use of different cost functions describes surface-level ozone ($O_3$) concentrations for the Marienfelde-Schichauweg monitoring station in urban Berlin. The monitoring technique used for $O_3$ is UV absorption. An exceedence is defined as an occurrence of the daily maximum of the 8 hour moving average exceeding a threshold concentration of $120\mu g/m^3$. A variety of variables are used as inputs into the models, and are listed in table 1. Using $O_3$ data from previous days allows the persistence of $O_3$ in the atmosphere to be taken into account. The temporal inputs allow for general patterns in time, in particular "day of the week" can be considered a weak surrogate variable for precursor traffic emissions data, and "time index" allows for non-stationary modelling of the data. The choice of meteorological variables used was guided by expert knowledge of the domain from the provider of the datset (Schlink, 2001). Note the the use of a Bayesian regularisation scheme with a Laplace prior in training the multi-layer perceptron models means that redundant inputs are likely to be suppressed if not entirely pruned from the network (Williams, 1995).

The data originally existed in the form of hourly readings, covering the period 1997-2000. Five models were evaluated: generalised linear models using logistic regression and homoscedastic Gaussian regression cost functions (denoted by GLM-LR and GLM-HoGR respectively) and multi-layer perceptron models using logistic regression, homoscedastic Gaussian regression and heteroscedastic Gaussain regression cost functions (MLP-LR, MLP-HoGR and MLP-HeGR respectively). A simple cross validation procedure was used in order to obtain a robust estimate of true generalisation ability of each model (Stone, 1974). The available data were divided into segments according to year. Four identical models were then trained using different permutations of three of the four segments of data and tested on the unused segment. The test statistics quoted in this section are then the arithmetic means of those

statistics evaluated over the test segment for each model.

A plot of the true daily time series is given in figure 2 along with the fitted predictions from the GLM-HoGR (simplest) and MLP-HeGR (most sophisticated) models. In order to avoid over-complicating the figure the outputs of the other models are not included, and only the second 100 days of the year 2000 are given, containing 27 of the 31 observed exceedences for that year. Both the MLP-HeGR and the GLM-HoGR fit the observed data well, with more than half of the exceedences correctly identified.

### *4.1 Log-likelihood Analysis*

Usually the global fit of a model is formally measured by using the root mean square error (RMSE) or equivalently the sum-of-squared error (SSE) or even the mean absolute error (MAE). These metrics provide an indication of the agreement between the true time series and the mode of the predictive distribution, but does not provide an indication of the suitability of the statistical assumptions made, such as whether the data is heteroscedastic in nature. The global performance measure used here is the *log-likelihood*, as this allows a fair and direct comparison of models based on different distributional assumptions on a common scale. As the homo- and heteroscedastic Gaussian regression models also provide an estimate of the probability of an exceedence, the log-likelihood for both the occurrence of an exceedence and the predicted pollutant concentration can be computed for the GLM-HoGR, MLP-HoGR and MLP-HeGR models. For either the prediction or classification problem, direct comparison of the log-likelihoods is possible: the model with the higher log-likelihood is "more likely" to have generated the observed data, and therefore can be considered superior assuming equal model complexity. To make allowances in comparing simple models to more complicated models, the Akaike Information Criterion (AIC) (Akaike, 1973) is often used, which penalises the log-likelihood of models with a greater number of parameters. However, this

method depends on all models being trained to a unique minimum of the cost function (the negative log-likelihood), which is not appropriate for the multi-layer perceptron networks considered here, as the training algorithm can only be guaranteed to converge to a local minimum of the cost function. The log-likelihood statistics for the $O_3$ dataset are given in table 2.

For the task of predicting exceedences of the statutory threshold, the logistic regression models out-perform the other models. The performance of the MLP-HeGR model, being the most sophisticated, is somewhat disappointing. This can be explained by recalling that maximum likelihood estimates of variance-like quantities are known to be biased (Bishop, 1995). If the model of the conditional mean over-fits the data, the apparent variance of the noise process will be reduced, and so the predicted variance will often be optimistically small. Hence, on rare occasions an overly-confident prediction is made (i.e. with very small variance) which fits poorly with the observed data, disproportionately deflating the log-likelihood for both the classification and the prediction problems. If we consider only the predicted mean from the MLP-HeGR model, and estimate the (unbiased but homoscedastic) variance using the formula:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2,$$

we get the adjusted log-likelihood values shown in parentheses in table 2. The comparison of log-likelihoods now corresponds to a comparison of the root-mean-squared-error for the task of predicting pollutant concentrations. The MLP-HeGR model now compares well with the logistic regression models for the classification problem, and is even more favourably compared with the other Gaussian regression models for the concentration prediction task. This suggests that the MLP-HeGR model is superior in predicting the conditional mean of the distribution of pollutant concentrations, due to the assumption of a heteroscedastic noise process, although some procedure for avoiding overly precise predictions may be beneficial.

It is interesting to note the similarity in performance of the GLM-HoGR and

the MLP-HoGR model, and also the MLP-LR and GLM-LR models. This suggests that the data is essentially linear in structure with perhaps only a weak non-linear component. This agrees with a preliminary non-linearity detection study of the $O_3$ data (Foxall et al., 2001), although the physics of $O_3$ in the atmosphere (Nunnari et al., 1998) suggests non-linear models would be more appropriate.

### *4.2 Exceedence Summary*

Specific health risks associated with air pollution can occur when key air quality thresholds are exceeded, and so in air pollution modelling it can be argued that it is most important to accurately predict these exceedences. Table 3 gives a summary of the 8 hour average exceedence predicted by each model. There were 108 exceedences in the $O_3$ dataset covering the period 1997–2000. The definition of the "best" classifier depends upon the ratio of false-positive to false-negative costs deemed appropriate for a given end-user application. In this instance, relative importance of ensuring the prediction of exceedences or the avoidance of false alarms. This approach allows individual air quality managers to optimise this trade-off to suit their own area of interest. Given a $50 : 50$ ratio of misclassification costs, the MLP-HoGR model correctly predicts the most exceedences (72) but at the price of an increased number of false alarms (27), although the GLM-HoGR model has a higher false alarm percentage. The MLP-HeGR model has the fewest false alarms and a "competitive" number of correctly predicted exceedences, and so is likely to be the preferred classifier for a range of misclassification costs.

### *4.3 McNemar's Test*

Given two classifiers $A$ and $B$, which classify each test pattern either correctly or incorrectly, McNemar's test (McNemar, 1947) decides whether the propor-

tion of times that $A$ is correct and $B$ is incorrect is essentially the same as the proportion of patterns for which $A$ is incorrect and $B$ is correct. Table 4 gives the probabilities of the paired classifiers being essentially the same for each of the possible pairings. The lower triangle of the table indicates the better classifier according to this system, measured by the total number of misclassifications. Table 4 shows here that the MLP-HeGR model is superior to all other models, although none of the model differences are even close to being statistically significant. All of the non-linear models outperform the two linear models. Clearly McNemar's test in its current form is not particularly informative unless it is known that there are equal misclassification costs for false-negative and false-positive errors.

### *4.4 ROC Curves*

For most "interesting" classification problems, it is difficult if not impossible to find a classifier which will classify every future pattern correctly. One therefore has to choose a classifier which keeps the misclassifications to a minimum. Of course, there are two different possible types of misclassification – false-positives and false-negatives, and often the act of altering a classification system so that one type of error is reduced increases the probability of the other type of error. The *Receiver Operating Characteristic* (ROC) curve graphically displays the trade-off between false-negative (1− true-positive) and false-positive rates obtained by varying the classification criteria – in this case, the probability threshold at which an exceedence is predicted (which depends on the most appropriate ratio of misclassification costs). The area under the ROC curve gives the effectiveness of the classifier, assuming nothing is known about the optimal operational ratio of misclassification costs, the closer the area to unity the better the classifier (Bradley, 1997; Adams and Hand, 1999). Table 5 gives the areas under the ROC curves for each model. Here the MLP-HeGR model again performs best, just ahead of the two logistic regression

models, the two homoscedastic Gaussian regression techniques providing the worst performance.

## 5 A Web-based Demonstrator

The key focus of the research presented here is on the construction of statistical models of air quality datasets that can be used to predict episodes of poor air quality, such that the end-users' estimates of the true misclassification costs are taken into account. Unfortunately it is often difficult for end users to express the relative importance of false-positive and false-negative misclassification errors numerically. We have therefore constructed a web-based demonstrator system for our best model (MLP-HeGR), illustrating the effects of changing the ratio of misclassification costs on the predictions made by a statistical classifier. This allows a potential end-user to determine the impact on their own sphere of interest of predictions made using different sets of misclassification costs. The demonstrator is provided in the form of a Java applet, shown in figure 3, and is accessible via the URL `http::/theoval.sys.uea.ac.uk/~gcc/projects/appetise/Berlin_S02.html` . The demonstrator incorporates a model used to predict the concentration of sulphur dioxide in urban Berlin, using an MLP-HeGR model (demonstrators for other pollutants, including $O_3$, are under development). The top panel displays the probability of an exceedence predicted by the model, the lower panel displays the observed $S0_2$ concentration and the conditional mean and $\pm 1$ standard deviation error bars, the horizontal line giving the statutory threshold concentration. The slider bar at the bottom of the applet allows the user to vary the ratio of misclassification costs, which affects the number of false-positive and false-negative errors made by the model. We hope that this will prove a useful tool for prospective end-users, helping them to consider quantitatively the effects of misclassification costs for their own application.

## 6 Summary

The multi-layer perceptron heteroscedastic Gaussian regression (MLP-HeGR) model provides the best estimate of the conditional mean of the concentration of $O_3$ and the best predictor of exceedences regardless of misclassification costs of the models investigated. It should be noted that the models described in detail here here contributed a rigorous intercomparison of a wide variety of statistical models for the task of predicting ground level ozone in a range of European locations, and were found to be competitive (Schlink et al., 2002). Unfortunately the variance estimated by this model is sometimes optimistically low, leading to an inflated likelihood statistic. However it is clearly worthwhile pursuing models incorporating more complex distributional assumptions as this may lead to both more accurate models for forecast purposes, but also a greater understanding and quantification of the physical processes giving rise to poor air quality (e.g. a strong indication that in this case the noise process is indeed heteroscedastic).

A key advantage of the Gaussian regression models is that they can be doubly calibrated; not only is it possible to determine the probability that a pollutant exceeds a fixed threshold, these models can still be used following a change in threshold level, due perhaps to the introduction of more stringent legislation. However, the assumption of constant variance would appear to be inappropriate for the datasets considered, leading to the two homoscedastic Gaussian regression techniques (GLM-HoGR and MLP-HoGR), providing poor predictions of the probability of an exceedence. When misclassification costs are not equal these two models will provide a relatively poor warning system. The MLP-HeGR model however, with the more flexible heteroscedastic variance model, can give both accurate predictions of exceedence probabilities and accurate predictions of pollutant concentrations, and so is to be preferred over the other models considered.

## References


Adams, N. M. and Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32:1139–1147.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *2nd International Symposium on Information Theory*, pages 267–281, Armenia, USSR. Tsahkadsov.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

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

Foxall, R. J., Krcmar, I. R., Dorling, S. R., Cawley, G. C., and Mandic, D. P. (2001). Nonlinear modelling of air pollution time series. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2001)*, volume VI, pages 3505–3508, Salt Lake City, Utah.

Gardner, M. W. and 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. and Dorling, S. R. (1999a). Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. *Atmospheric Environment*, 33(5):709–719.

Gardner, M. W. and Dorling, S. R. (1999b). Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment*, 34(1):21–34.

Grieg, A. J. e. (2000). Air pollution episodes : modelling tools for improved smog management (APPETISE). In *Proceedings of the Eighth International Conference on Air Pollution*, pages 89–98, New Hall, Cambridge University, U.K.

Hassibi, B. and Stork, D. G. (1993). Second order derivatives for network pruning: optimal brain surgeon. In Hanson, S., Cowan, J., and Giles, C., editors, *Advances in Neural Information Processing Systems*, volume 5, pages 164–171, San Mateo, CA. Morgan Kaufmann.

Hopfield, J. J. (1987). Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proceedings of the National Academy of Sciences*, 84:8429–8433.

McNemar, I. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384.

Nix, D. A. and Weigand, A. S. (1995). Learning local error bars for nonlinear

regression. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing*, volume 7, pages 489–496. MIT Press.

Nunnari, G., Nucifora, A. F. M., and Randieri, C. (1998). The application of neural techniques to the modelling of time-series of atmospheric pollution data. *Ecological Modelling*, 111:187–205.

Schlink, U. (2001). personal communication.

Schlink, U., Dorling, S., Pelikan, E., Nunnari, G. Cawley, G., Junninen, H., Grieg, A., Foxall, R., Eben, K., Chatterton, T., Vondracek, J., Richter, M., Dostal, M., Bertucco, L., Kolehmainen, M., and Doyle, M. (2002). Rigorous inter-comparison of ground-level ozone predictions. *Atmospheric Environment* (submitted).

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B*, 36(1):111–147.

Weisberg, S. (1985). *Applied linear regression*. John Wiley and Sons, New York, second edition.

Williams, P. M. (1995). Bayesian regularisation and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143.

Williams, P. M. (1996). Using neural networks to model conditional multivariate densities. *Neural Computation*, 8:843–854.

Table 1
Input variables. Round brackets $(\cdot)$ refer to time delay (in days) of input relative to the prediction horizon, the numbers in square brackets $[\cdot]$ represent the hourly data used to form averages etc. Defaults are current day $(t-0)$ day and all hours $[1-24]$ respectively.

| **Category** | **Input variable** | |
|---|---|---|
| **Pollutant** | mean $O_3(t-2)$ | max $O_3(t-2)$ |
| | mean $O_{3[1-18]}(t-1)$ | max $O_{3[1-18]}(t-1)$ |
| | mean $NO_{x[1-18]}(t-1)$ | max $NO_{x[1-18]}(t-1)$ |
| **Temporal** | sine of Julian Day | cosine of Julian Day |
| | day index | day of the week |
| **Meteorological** | mean temperature | max temperature |
| | mean global radiation | mean humidity |
| | mean wind $\text{speed}_{[7-10]}$ | mean wind $\text{speed}_{[15-19]}$ |
| | mean wind direction | mean radiation balance |

Table 2
Log-likelihood for $O_3$ models, numbers in brackets $(\cdot)$ refer to values obtained using a homoscedastic variance structure with the MLP-HeGR model (see text for details).

| **Model** | **log-likelihood for occurrence** | **Rank** | **log-likelihood for prediction** | **Rank** |
|---|---|---|---|---|
| **GLM-LR** | -155.73 | 2 (3) | N/A | N/A |
| **MLP-LR** | -154.88 | 1 (1) | N/A | N/A |
| **GLM-HoGR** | -169.30 | 4 (4) | -5807.91 | 3 (3) |
| **MLP-HoGR** | -172.61 | 5 (5) | -5748.93 | 2 (2) |
| **MLP-HeGR** | -158.37 (-155.01) | 3 (2) | -5672.10 (-5619.90) | 1 (1) |

Table 3
Summary of the prediction of threshold exceedences by model type.

| **Model** | **observed exceedences** | **predicted exceedences** | **number correctly predicted** | **number of false alarms** |
|---|---|---|---|---|
| **GLM-LR** | 108 | 90 | 66 (61.1%) | 24 (26.7%) |
| **MLP-LR** | 108 | 91 | 68 (63.0%) | 23 (25.3%) |
| **GLM-HoGR** | 108 | 85 | 60 (55.6%) | 25 (29.4%) |
| **MLP-HoGR** | 108 | 99 | 72 (66.7%) | 27 (27.3%) |
| **MLP-HeGR** | 108 | 86 | 66 (61.1%) | 20 (23.3%) |

Table 4
Outcome of McNemar's test for the statistical significance of differences in the performance of classification models.

| **Model** | **GLM-LR** | **MLP-LR** | **GLM-HoGR** | **MLP-HoGR** | **MLP-HeGR** |
|---|---|---|---|---|---|
| **GLM-LR** | 1 | 0.7003 | 0.4008 | 0.7423 | 0.5563 |
| **MLP-LR** | MLP-LR | 1 | 0.1547 | 1.0000 | 1.0000 |
| **GLM-HoGR** | GLM-LR | MLP-LR | 1 | 0.1748 | 0.1273 |
| **MLP-HoGR** | MLP-HoGR | tie | MLP-HoGR | 1 | 1.0000 |
| **MLP-HeGR** | MLP-HeGR | MLP-HeGR | MLP-HeGR | MLP-HeGR | 1 |

Table 5
Area under the ROC curve for $O_3$ prediction models.

| Model | Area under ROC | Rank |
|---|---|---|
| **GLP-LR** | 0.9741 | 2 |
| **MLP-LR** | 0.9724 | 3 |
| **GLM-HoGR** | 0.9661 | 5 |
| **MLP-HoGR** | 0.9704 | 4 |
| **MLP-HeGR** | 0.9762 | 1 |

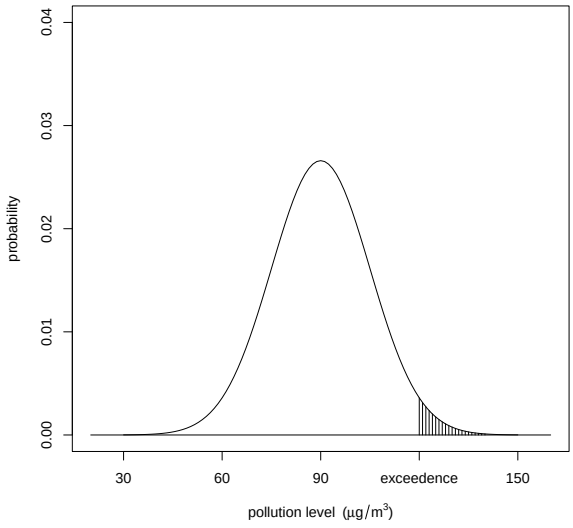


Fig. 1. Predictive distribution given by a Gaussain regression model, the shaded area represents the probability of an exceedence of a fixed threshold.

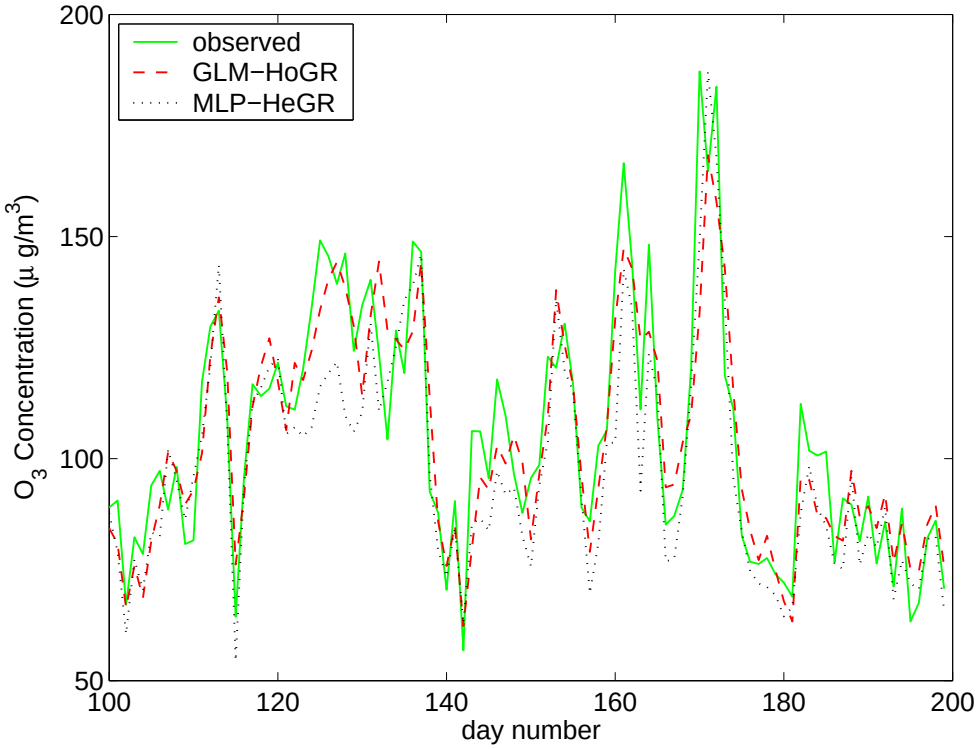


Fig. 2. True time-series and predicted concentrations given by MLP-HeGR and GLM-HoGR models for the second 100 days of 2000 for the Berlin surface-level $O_3$ dataset.

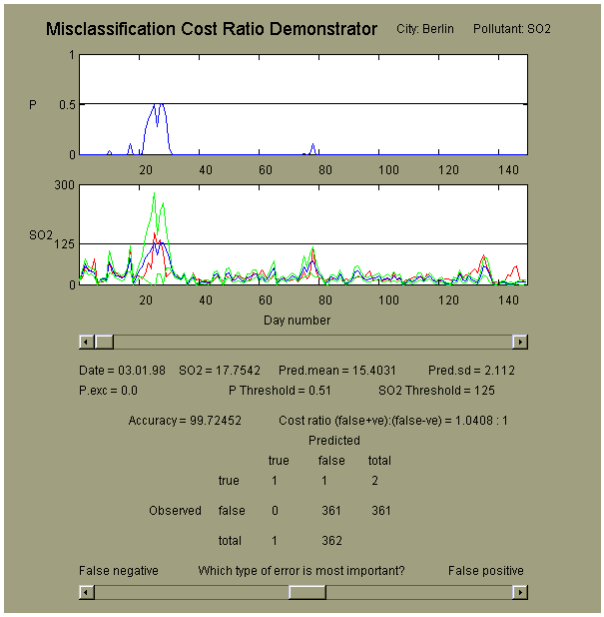


Fig. 3. Java applet demonstrating the trade-off between false-positive and false-negative misclassification costs (`http::/theoval.sys.uea.ac.uk/~gcc/projects/appetise/Berlin_S02.html` ).