ANALYSIS OF CHEMICAL EXPOSURE THROUGH INHALATION USING HYBRID NEURAL NETWORK

T. RAJKUMAR and HANS W. GUESGEN
Department of Computer Science, University of Auckland
New Zealand

ABSTRACT

In this analysis, human health risk through inhalation due to exposure to Benzene from vehicular emissions in New Zealand is assessed as an example of the application of a hybrid neural network. Exposure factors affecting the inhalation are inhaled contaminant, age, body weight, health status and activity patterns of humans. There are four major variables affecting the inhaled contaminant viz., gas emissions from motor vehicles on the road, wind speed, temperature and atmospheric stability. The topic of uncertainty applies equally to all variables involved in exposure analysis. Neural network and fuzzy theory is implemented to solve the uncertainty, which exists to a greater extent. The architecture of hybrid neural network that is used to estimate the exposure of carcinogens through inhalation is explained in detail in this paper.

1. INTRODUCTION

Humans are exposed daily to low concentrations of many natural and man-made chemical substances. Although many of these substances can be toxic at high concentrations, typical exposures are far below the effect concentrations. Our society seems to be concerned about the health risks associated only with the synthetic chemicals, regardless of their proportional contribution to the total agonist activity. The adverse effects caused by the toxic chemicals made the scientific community realise that there is a need for developing sound approaches for risk assessment. Risk assessment includes one or more of the following components: Hazard identification, dose response assessment, exposure assessment and risk characterization. Exposure assessment is a central component of the quantitative risk assessment procedures, and the process of estimating the environment concentration of a contaminant plays a significant role.

The inherent variability in environmental and exposure related parameters are a major source of uncertainty. For example, meteorological and hydrological conditions change seasonally at a given location, soil characteristics exhibit large spatial variability, and human activity patterns depend on the age, sex and geographic location of specific individuals in the population. Existing exposure models are not reliable due to the associated uncertainty with variables. A hybrid neural network approach proved more advantageous over traditional phenomenological or semi-empirical formula and it is able to develop a mapping of the input and output variables. Such a mapping can be used to estimate the output variables as a function of input variables. In this paper, the hybrid neural network incorporates a modified back propagation neural network and fuzzy logic to estimate the inhaled chemical concentration.

There has been growing concern about the effects of human exposure to various toxic and hazardous substances in the environment. Once dose-response data are available, it is vital to assess how this applies to the general population, especially that community living close to sources. The model developed here employs the latest techniques to provide quantitative exposure data for air toxic, and demonstrates this using a measured data set. The ultimate aim of this research is to produce a general model of toxic exposures, which can be applied to any community, for any substance where the dose-response is known.

2. INHALATION EXPOSURE

The health risk associated with human exposure to airborne toxic is a function of concentration of air pollutants, chemical species, duration of exposure, and inhalation rate. The concentration of air pollutant is derived as a function of source of emissions, wind speed, ambient temperature and atmospheric stability. The inhalation rate is affected by numerous individual characteristics, including age, body weight, health status and levels of activity. The United States Environmental Protection Agency adopts the following general equation for calculating average daily dose (ADD) for inhalation exposure:

$$ADD = \frac{C \times IR \times ED}{BW \times AT}$$

C = Contaminant concentration in inhaled air ($\mu g/m^3$)
IR = Inhalation rate (m^3/day)
ED = Exposure duration (days)
BW = Body Weight (kg) and
AT = Averaging time (days) (life time)

Even in situations where actual exposure-related measurements exist for the above variables in the equation, assumptions or inferences will still be required because data are not likely to be available for all aspects of the exposure assessment. Moreover, the data are available may be of questionable or unknown quality. The neural network may reduce the noise of the data by learning the patterns. The imprecise input information for the exposure model is expressed as fuzzy sets using linguistic variables such as low, medium and high. The fuzzy and neural networks coupled together provides a better result for solving the uncertainty than the conventional methods and it is explained in detail in the following sections.

3. HYBRID NEURAL NETWORK

The term hybrid refers to systems that consist of two or more integrated subsystems, each of which may have a different representation and inference technique. The subsystems are assumed to be tied together semantically and influence each other. In this paper, neural networks and fuzzy logic are the two subsystems and the architecture of the system is explained in the figure 1.
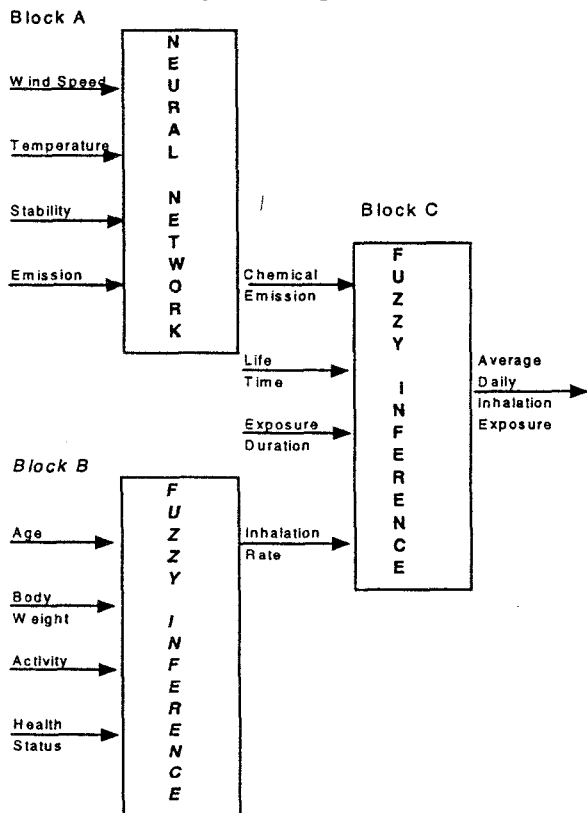

FIG 1 HYBRID NEURAL NETWORK ARCHITECTURE

Each block is represented by a subsytem and coupled together to derive the chemical absorption through inhalation. Block A consists of a neural network and is designed using a modified back propagation algorithm. Block B and Block C use a set of fuzzy rules and derive the average chemical exposure through inhalation based on fuzzy inference. In the following sections details about the blocks will be explained.

Modified Backpropagation Algorithm (Block A)

A neural network is conceptually comprised of a collection of nodes and connections. The basic elements of a network are called neurons; they represent the sites that process information. The interconnecting links between the processing units, or neurons, are called synapses. Each synapse can be characterized by a weight, which is represented by a numerical value. All neurons in the adjacent layers are connected and the flow of information is restricted to the forward direction. The network consists of three layers: input layer, hidden layer and output layer. The hidden layer enables the network to learn relationships between input-output variables through suitable mappings. Among the many neural network models, the backpropagation algorithm is one of the better known and frequently used.

A backpropagation network typically starts out with random set of connection weights. The network adjust its weights based on the delta learning rule each time it sees a pair of input-output patterns. Each pair of patterns goes through two stages of activation: a forward pass and a backward pass. The forward pass involves presenting a sample input to the network and letting the activations flow until they reach the output layer. During the backward pass, the network's actual output (from the forward pass) is compared with the target output and errors are computed for the output units. The weights connected to the output units can be adjusted in order to reduce those errors. The error estimates of the output units are then used to derive error estimates for the units in the hidden layers. Finally, errors are propagated back to the connections stemming from the input units. After each round of forward-backward passes, the system "learns" incrementally from the input-output pair and reduces the difference (error) between the network's predicted output and the actual output. After extensive training, the network will eventually establish the input-output relationships through the adjusted weights of the network.

To improve some aspect of the learning, such as speed of training and learn with consistency, or generalization, the new backpropagation algorithm performs two passes, a pattern pass and a batch pass. The presentation of data sets varies in the two passes.

Pattern Pass: In this pass, each input-output pair is submitted to the network for learning. Training

continues until each pattern satisfies the error of tolerance.

i. Create the architecture of the neural network with one bias node in the input layer, which always has the value 1.
ii. Assign random values between -1 and +1 to the weights between the input and hidden layers and the weights between the hidden and output layers.
iii. Each connection in the architecture will have differential learning and momentum rates. The learning rates and momentum rates are assigned randomly from 0.9-0.5 and 0.5-0.1
iv. Based on the continuous and differentiable sigmoid activation function F(x) as shown below, the network computes the node weights from the input layer to the next hidden layer. It continues this forward activation process until the node weights at the output layer are computed.

$$F(x) = \frac{1}{1 + e^{-x}}$$

v. Working backward from the output layer to the input layer, errors of the units in each layer are computed. An error correction formula (delta rule) is adopted which is based on the derivative of the activation function and which tries to minimize the error of the output units. The error function defines a surface over weight space, and the weights are modified in the direction of the gradient of the surface (gradient descent). For the output layer, error computation is based on the difference between the target (actual) output y and the network computed output o as follows:

$$\delta = o(1-o)(y-o)$$

For other layers, the error is computed as follows:

$$\delta = o(1-o)\sum_{i=1}^{n} \delta \ w$$

vi. The errors are used to adjust the connection weights in the network. A learning rate (β) and momentum factor (α) can be used to improve the learning process. The new connection weight is obtained using the formula

$$w_{t+1} = w_t - \delta \ \beta \ o \ + \alpha \ w_t$$

Where o, δ are measured at time t+1 and w_t is the weight change experienced during the previous backward-forward pass (time t). Each and every connection in the network has its own learning and momentum rates.

Batch Pass: After the training is completed in the pattern pass, the network is tested for its convergence of the training data sets. In our experiments, 90-95 percent of the training data set reached the convergence and 5 percent of data did not reach convergence for the required error of tolerance. In the batch pass, the whole training data set is presented to the network similar to how it is done during standard backpropagation.

There are two differences in the algorithm between the pattern pass and the batch pass.

i. The learning rate is constant (β=0.8) and momentum rate is considered similar to pattern pass.
ii. The data sets are presented as a whole batch and complete round of forward and backward passes are performed to achieve the mapping of input-output data sets.

In this modified algorithm, the pattern pass performs the stabilization and generalization of the network. The batch pass tunes the mapping to a greater accuracy. The number of cycles for training is less than the standard backpropgation algorithm and it is applied to determine the chemical exposure of benzene emissions from motor vehicles as an example.

Fuzzy Inference (Block B and Block C)

The use of fuzzy set theory will allow the user to include the unavoidable imprecision of the data. Fuzzy inference is the actual process of mapping from a given input to an output using fuzzy logic. The first step is to fuzzify the input variables and output variables against the linguistic sets (low, medium and high). Rules will be formulated based on the experts knowledge or on the basis of literature available. The input is always a crisp numerical value limited to the function of discourse of the input variable. If the antecedent of the rule has more than one part, the fuzzy operator is applied to obtain one number that represents the result of the antecedent for that rule. The logical AND is defined as the minimum of the fuzzy numbers and OR as the maximum. The implication method is defined as the shaping of the consequent based on the antecedent, and the output is a fuzzy set.

Aggregation is done through joining the output of the rules by a thread. The input for the defuzzification process is a fuzzy set and the output is the daily average inhalation of chemical. The centroid method is adopted for finding the centre of the area under the curve, which represents the daily average inhalation of chemical. The input variables for the block C arise from the output of block A and block B. The block A estimates the final chemical concentration which is being inhaled by humans after all degradation process. Block B calculates the inhalation rate of humans. Integration between all blocks is loosely coupled since neural network training is slow. The neural network component should be preprocessed and is kept in a separate file. The output of Block A is fuzzified into three membership functions based on the statistical significance.

4. EXPERIMENTAL SETUP

Definition of Variables

Neural Networks (Block A): The success of the neural network greatly depends on defining all of the appropriate influencing variables. In this paper, the following variables are selected for determining the exposure: gases emissions from motor vehicles on the road, wind speed, atmospheric stability and ambient temperature. The release height, sampling height and terrain, however, is not varied during the experiment. Hence the variables listed above are believed sufficient to completely characterize the dispersion and exposure of benzene. The data pertaining to 1990 is considered for training and 1991 for testing. Due to the non-availability of continuous data for benzene monitoring, we have established a correlation between NO_x and Benzene from 6 months data using the model shown in Figure 2.
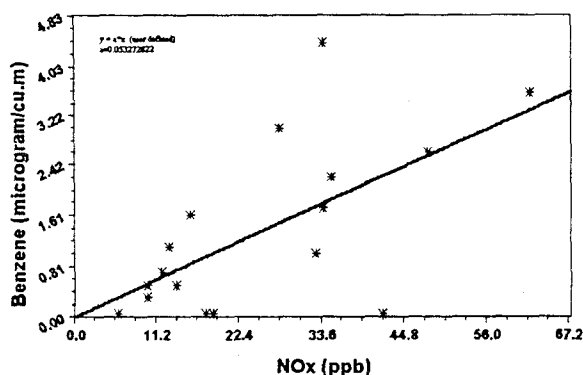
Benzene and NOx at Christchurch City

FIG 2 RELATIONSHIP BETWEEN BENZENE AND NO_x

From the correlation, the output of the neural network is established. The gases emissions from motor vehicles is derived as

$$\text{Input Concentration} = \frac{N}{M} \times 30 + 0.9 \times C$$

N = No. of vehicles passed through the road at that hour
M = Maximum No. of vehicles passed in the week
C = The concentration of benzene measured at that hour

If there are no vehicles on the road, the input concentration will be calculated as 50% of the previous hour input concentration of benzene and measured benzene in that hour. The value 30 is our assumed value of the maximum concentration of benzene, which occurs at the site, for peak traffic. It is derived from examining the long-term monitored data set for worst case conditions. The benzene in the air exists predominantly in the vapor phase, with residence times varying between a few hours and a few days, depending on the environment and climate, and on the concentration of

hydroxyl radicals, as well as nitrogen and sulfur dioxides.

The ranges of the input variables must be defined clearly before the neural network training and testing. After training the neural network is used to predict the out-of-range data, it will lead to a misleading result. The training data has been well spread over the entire range. Although every attempt is made to investigate the entire range of Turner's Stability class (1-7), this proved to be difficult because of the limited variety in the datasets. The range of input and output variables are defined for the training set in Table 1:

Variable	Range		Unit
Emissions from Motor vehicles	0.128	69.997	$\mu g/m^3$
Wind Speed	0	21.3	M/sec
Temperature	-2.65	35.2	°C
Atmospheric Stability	1	7	Turners
Output Concentration (Exposure)	0	65.552	$\mu g/m^3$

TABLE 1 RANGES OF VARIABLES OF TRAINING (Block A)

Fuzzy Inference (Block B): The input variables for block B are the age, body weight, health status and activity patterns. Block B determines the inhalation rate of the humans. The inhalation rate is divided into three sets viz., adults, children and high activity persons. Presently the research is focussed to only adults and it will be extended all groups of people. Table 2 describes the input variables and the range of values that have been chosen.

Variables	Range	Distribution
Age (yr.)	18 - 80	triangular
Body Weight (kg)	45-110	normal
Health Status	0-1	normal
Activity (l/min)	23-64	normal
Inhalation rate (l/min.)	16-49	normal

TABLE 2 RANGES OF VARIABLES (Block B)

Fuzzy Inference (Block C): The input variables for the block C are the outputs from block A and B. The output from block A will be clustered into three linguistic variables, based on statistical inference the data. Apart from the two variables, exposure duration and lifetime is also forming the input variables. The output of the block C based on the rules derives the average daily inhalation exposure. The average daily inhalation exposure with dose response is used estimate the risk of particular chemicals.

314

5. NETWORK TRAINING

Having created data sets for training and testing, the next task is to choose an appropriate neural network architecture. The choice of a three layer neural network required the appropriate specification of a number of nodes in the input, hidden and output layers. The four input variables together with the bias are represented by five input neurons. One output neuron is used to estimate the exposure. The number of hidden neurons is set to be equal numbers of input neurons. The inputs to the neural networks are scaled values between 0.1 and 0.9. The scaling is performed using the following equation

$$A = r(V - V_{min}) + A_{min} \qquad r = \frac{A_{max} - A_{min}}{V_{max} - V_{min}}$$

V-Observed variable   A-Presentable variable

The program is executed by specifying the number of nodes in each layer, as well as specifying the value of the error tolerance. Every month data are presented to the network and an extensive training is performed. At Christchurch City, the four seasons are distinct and therefore the generalization for the entire season was not appropriate.

Network Testing

The trained network is used to make predictions of emission rates using data, which it had not encountered during its training. The ranges of variables in the test data all fell within the range used to train the neural network. A sample result of prediction using the network is presented in Figure 3, which compare the emission rates predicted by the neural network with the measured exposure.

Benzene Prediction (June 1991)
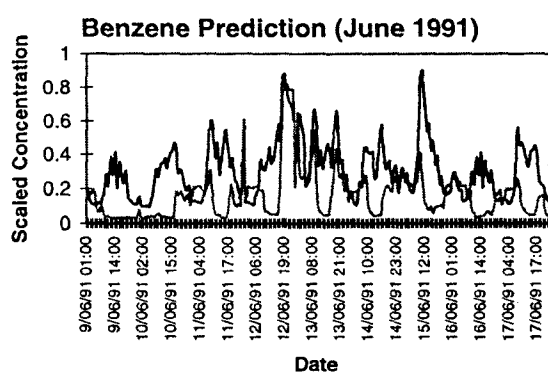
——— Measured Concentration

——— Neural Network prediction

FIG 3 PREDICTION OF BENZENE USING MODIFIED NEURAL NETWORK

Figure 3 shows that the predictive model is reasonably stable and replicate the important features of the data record. For instance the peak values are modeled, although not all peaks are explicitly resolved. The morning increases are resolved reasonably well.

There are several factors, which partially explain the variance between actual and modeled data. The first relates to the use of NO_x concentrations as a surrogate for benzene. As can be seen in Figure 2, the relationship has a considerably degree of scatter which will be reflected in the modeling. Secondly, the roadside concentrations depend very strongly on complex meteorological conditions. Our model uses the main factors, but does not incorporate all possible factors. The encouraging results suggest that the dominant factors have been used, but others may need to be used in more sophisticated approaches. Thirdly, the analysis and modeling has been based on averaging periods of one hour, as a practical working time period. However the true situation changes substantially on shorter periods, perhaps down to the order of a few minutes. Finally, the source data, which arises from traffic volumes, has been estimated and assumed as mainly coming from one road. This will not in generally be true, but is probably an adequate assumption for this study. However despite the level of variance, the model is capable of showing the extent and incidence of high concentrations, and it is these that dominate the exposure relationship, which is ultimate goal of this research.

The network was tested against the standard back propagation algorithm and the modified backpropagation algorithm; the result is shown in the following Table 3:

Month	No. of Data sets	Cycles	
		Standard	Modified
Jan	720	2	2
Feb	672	2	2
Mar	507	2	2
Apr	719	268	135
May	711	8880	7084
Jun	691	10616	7699
Jul	742	6893	4341
Aug	723	218	134
Sep	720	589	278
Oct	740	2	2
Nov	475	2	2
Dec	655	2	2

TABLE 3 NO. OF CYCLES FOR TRAINING

6. RESULTS AND CONCLUSION

The use of a neural network provides a new perspective on how to model complex, non-linear problems. The neural network approach works well only if it is provided with good examples from which it can extract patterns or identify suitable features of the training data. The data has to be sufficiently scattered over a range of variables. Most models of atmospheric dispersion are based on empirically estimated parameters. The main advantage of neural network approach is that it lessens the burden of finding a mathematical form appropriate for fitting various experimental data. The trained network makes almost instantaneous predictions due to a relatively simple computational procedure. However, most phenomenological models are computationally intensive, and in general do not reduce the uncertainty in providing reliable estimates. Traditional statistical models use established performance measures such as regression and correlation coefficients to determine the extent of fit of the data by a model. The neural networks, however, have no such established criteria. The parameter that is involved in this neural network design is the number of hidden nodes, the learning rates and the momentum constant.

Block B and C in Figure 1 are under development. This research has shown the excellent possibilities that exist for using fuzzy logic and neural networks as powerful tools for exposure analyses. Even these preliminary results on just one contaminant, offer encouragement for application to a wide range of other contaminants, in a number of scenarios. Further work will be needed using other datasets to confirm this applicability, and this is being undertaken as the next step in the research. Depending on the nature of the application and the strength of the internal data patterns the neural network learns quite well. This applies to problems where the relationships may be quite dynamic or non-linear. Neural networks provide an analytical alternative to conventional techniques, which are often limited by strict assumptions of normality, linearity, variable independence etc. Because a neural network can capture many kinds of relationships, it allows the user to quickly and relatively easily model phenomena which otherwise may have been very difficult or impossible to explain. This type of modeling proved to be highly successful in the application described in this paper.

7. REFERENCES

[1] United States EPA, *Exposure Factors HandBook,* 1996.
[2] United States EPA, *Risk Assessment Guidelines,* 1986

[3] Ministry of Health/Public Health Commission, *Survey of Benzene and Lead in Air,* August 1995.
[4] WHO, *Benzene,* 1993.
[5] T.S. Clarkson, R.J. Martin, J. Rudolph and B.W.L. Graham, Benzene and Toluene in New Zealand Air *Atmospheric Environment,* pp 569-577.
[6] M.A. Rege and R.W. Tock, "A simple neural network for estimating emission rates of hydrogen sulfide and Ammonia from Single Point Source", *Journal of the Air & Waste Management Association,* October 1996, pp 953-962
[7] C. Jondarr, *"Back propagation Family Technical Report C/TR96-05",* Macquarie University, 1996
[8] Young W Lee, Mohamed F dahab and Istavan Bogardi, "Nitrate Risk Assessment using Fuzzy set approach", *Journal of Environmental engineering,* 1996
[9] O.Wong, "An industry wide mortality study of chemical workers occupationally exposed to Benzene II Dose response analyses", *British Journal of Industrial Medicine,* 1987, pp 382-395
[10] Cox, Earl, *"The fuzzy systems handbook: a practitioner's guide to building and maintaining fuzzy systems",* Boston AP professional, 1994

ACKNOWLEDGMENT