Pork farm odour modelling using multiple-component multiple-factor analysis and neural networks

Kevin R. Janes\textsuperscript{a}, Simon X. Yang\textsuperscript{a,*,} Roger R. Hacker\textsuperscript{b}

\textsuperscript{a}School of Engineering, University of Guelph, Guelph, Ont., Canada N1G 2W1
\textsuperscript{b}Department of Animal and Poultry Science, University of Guelph, Guelph, Ont., Canada N1G 2W1

Received 29 September 2003; received in revised form 23 September 2004; accepted 19 October 2004

Abstract

The number of non-farming rural residents surrounding pork farm operations has increased greatly in many pork-producing nations. These non-farming rural residents consider pork farm odour emissions to be a serious health threat. This has led to significant lobbying to limit the expansion and development of pork operations, and consequently reduced economic growth in the pork industry. This has led pork producers to search for effective methods to mitigate the odourous emissions, and researchers to seek methods to adequately model the odour. Extensive research has been performed on the use of single-component analysis to model the odour. Several researchers have used multiple-component analysis to extend the effectiveness of previous models. Since odour generation factors, like temperature, also contribute to the odour, the next logical approach to modelling pork farm odour is multiple-component multiple-factor analysis. It is proposed that the multiple-component neural network model be extended to make use of multiple-component multiple-factor analysis. First, a neural network model and a linear multiple regression model are developed and compared using multiple-component analysis to demonstrate the better modelling technique for pork farm odour. The neural network model of the pork farm odour yielded more accurate and precise odour intensity predictions than the linear multiple regression models, indicating that neural networks are the better modelling technique for this application. Subsequently, a multiple-component multiple-factor neural network model was developed and compared with the multiple-component neural network. The multiple-component multiple-factor neural network model generated performance gains, indicating that this approach is relevant to modelling pork farm odour. It is hypothesized that the extension of the multiple-component multiple-factor analysis to include additional significant odour components and odour generation factors in the neural network model will further improve model performance.

Keywords: Odour emissions; Pork farms; Neural networks; Odour modelling; Multiple-component multiple-factor analysis

1. Introduction

Over past decades, agricultural researchers have been attempting to adequately model pork farm odour with little success. The purpose of these modelling
efforts has been to increase understanding of the odour itself and allow a convenient and practical method for measuring pork farm odour. This need to easily measure pork farm odour has been created by public relation problems regarding odours emanating from swine operations. These public relations issues are primarily with non-farming rural residents who are not accepting of farm odours [14]. Researchers initially attempted to determine a single-component of the odour that could act as an indicator for odour intensity [12,16]. Single-component analysis of the odour using statistical tools or more advanced neural networks is insufficient because pork farm odour is complex requiring at least a thorough multiple-component analysis [7]. Pork farm odour is a very complex system that has over 200 different components that interact with each other and a variety of factors that contribute to the generation of the odour, such as relative humidity [15,21]. Zahn et al. [18] of Iowa State University have had some success with multiple-component analysis of the odour using a nine-component statistical model. It is proposed that multiple-component multiple-factor analysis of the pork farm odour using intelligent modelling techniques, specifically neural networks, will yield more accurate odour intensity predictions.

Neural networks can be used for modelling the relationships that exist within a complex system [17]. Neural networks are trained using data that is representative of the system being modelled, and typically the neural network will provide predictions or classifications as the output [8,13]. Neural networks are capable of automatically learning the relationship that exists between the selected odour components, contributing factors and odour intensity. Neural networks do not need any previous knowledge of the system being studied to determine the relationship that exists between the inputs and outputs [5,20].

The first objective of this research was to demonstrate that multiple-component neural network models of the pork farm odour will provide better overall performance than corresponding statistical models of the odour. This will be achieved by comparing the performance of multiple-component neural network and statistical models using ammonia and hydrogen sulphide as the studied components. The second objective is to demonstrate that multiple-component multiple-factor analysis of the pork farm odour will improve performance of pork farm odour models over simply multiple-component analysis of the odour. This will be achieved by comparing a multiple-component neural network model against a multiple-component, multiple-factor neural network model.

Researchers from the University of Minnesota collected the data used to develop and test the models in this study. Each sample in the dataset contains the air-phase concentrations for ammonia and hydrogen sulphide, time of measurement, date of measurement, outside temperature and the odour intensity value generated by human assessors.

Section 2 of this paper presents the dataset, statistical analysis tools, statistical modelling technique and the neural network architectures used in this investigation. In Section 3, the results are presented and compared for the various models. A conclusion drawn from the analysis of results is made in Section 4.

2. Methods

This section discusses the dataset used for model development and testing, the statistical tools used to analyse the results, and outlines the specific modelling techniques used in this research.

2.1. Data sets

Researchers from the Department of Biosystems and Agricultural Engineering at the University of Minnesota provided the data used in this investigation. These researchers compiled this data from 26 pork operations in the state of Minnesota for a database to be used by agricultural researchers. All the measurements were taken from swine housing sources, such as barns. The dataset has the air-phase concentration for ammonia and hydrogen sulphide. These two compounds have been thoroughly investigated as potential single component indicators because of their odour characteristics and ease of measurement [2,4,11,12,16]. The time and date of measurement have been recorded for each sample and the outside temperature is available. The time of day is an easy measurement to record, but is not a physical measure. However, time when the measure-
ment is taken can generally represent such measures as farm activity, ventilation, temperature, humidity, and other physical measures. Patterns emerge as to humidity levels throughout the day and when specific farm activities occur, which are captured by the time of day. The month of year is also an aggregate measure that represents physical measures, such as the predominant weather conditions in a particular season.

The odour samples were collected using a vacuum box fitted with a pump to draw air into 10 L Tedlar bags. The odour dilution threshold was determined within 48 h of sampling using trained human assessors and a dynamic olfactometer.

There are a total of 131 samples in the dataset available for the development and testing of models. Approximately 20% of the samples, or 26 data points, were randomly selected for model testing and the remaining 80% of the samples, or 105 data points, were used for model development. In this investigation, three models have been developed using the dataset, one linear multiple regression model and two neural network models.

All data used for model development has been normalized on the interval [0 1]. The normalization involved finding the largest positive number in the dataset and then dividing all data points in the set by the largest positive number.

2.2. Statistical analysis tools

The results from this investigation have been analysed using mean absolute error, standard deviation, and coefficient of determination. The mean absolute error (\(E_a\)) measures the mean of the absolute difference between the observed and predicted values [20]. The \(E_a\) was applied as the principal measurement for evaluating the various models in this investigation. The performance objective for each model is to attain an \(E_a\) as close to zero as possible. In order to evaluate the precision of each model, the standard deviation of the mean absolute error (\(\sigma\)) is also presented. A model’s precision in predicting odour intensity is greater the lower the standard deviation. The standard deviation is defined as:

\[
\sigma = \sqrt{\frac{\sum_{i=1}^{N} (y - \bar{y})^2}{N - 1}}
\]

Both the \(E_a\) and \(\sigma\) are presented as percentages to allow easier interpretation and comparison of model results.

The coefficient of determination, \(R^2\), is the percent of the variation that can be explained by the regression equation or simply the square of the coefficient of correlation [9]. The coefficient of determination \(R^2\) ranges on the interval \([0 1]\), where 0 indicates that there is a random or non-linear relationship between the variables, a value of 1 for \(R^2\) indicates a linear relationship [1]. The coefficient of determination is calculated by the following equation [9]:

\[
R^2 = \left[ \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \right]^2
\]

The variables \(S_{xx}, S_{yy}\) and \(S_{xy}\) are defined as:

\[
S_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})^2
\]

\[
S_{yy} = \sum_{i=1}^{N} (y_i - \bar{y})^2
\]

\[
S_{xy} = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})
\]

where \(N\) is the total number of samples; \(x\) and \(y\) are the input and output variables, respectively; \(\bar{x}\) and \(\bar{y}\) are their mean values; and \(i\) is an index number. The coefficient of determination was used to determine the correlation between the predicted values generated by the models and observed values of the test set.

2.3. Statistical model

Pork farm odour researchers have traditionally used statistical techniques to model the odour using single component or multiple component analysis. In particular, pork farm odour researchers have made extensive use of linear statistical models for both single component and multiple component analysis of the odour. Thus, linear statistical models were chosen in this investigation because they represent the state-of-the-industry. The limitation of this approach has been that other factors such as the temperature and month of the year cannot be considered in the statistical model. Linear multiple regression was selected as the statistical tool for modelling the data
in this study for comparison against neural network models for the odour. Multiple regression analysis deals with three or more variables and the relationship can be linear or non-linear [9]. The linear multiple regression equation is given as:

\[ y = b_0 + b_1 x_1 + \cdots + b_r x_r \]  

where \( x_1, \ldots, x_r \) are the component air-phase concentrations, \( y \) is the odour intensity, and \( b_0, b_1, \ldots, b_r \) are coefficients.

There are a set of equations that were used to determine the constants \( b_0, b_1, \ldots, \) for the equation, where \( r \) is 2 [9]:

\[ \sum_{i=1}^{N} y_i = Nb_0 + b_1 \sum_{i=1}^{N} x_{i1} + b_2 \sum_{i=1}^{N} x_{i2} \]  

\[ \sum_{i=1}^{N} x_{i1}y_i = b_0 \sum_{i=1}^{N} x_{i1} + b_1 \sum_{i=1}^{N} x_{i1}^2 + \sum_{i=1}^{N} x_{i1}x_{i2} \]  

\[ \sum_{i=1}^{N} x_{i2}y_i = b_0 \sum_{i=1}^{N} x_{i2} + b_1 \sum_{i=1}^{N} x_{i1}x_{i2} + b_2 \sum_{i=1}^{N} x_{i2}^2 \]  

The input variables \( x_1 \) and \( x_2 \) for these equations are the air-phase concentration of ammonia and hydrogen sulphide, respectively. The hydrogen sulphide is measured in micrograms per second per meter squared (\( \mu g \text{ s}^{-1} \text{ m}^{-2} \)) and the ammonia is measured in milligrams per second per meter squared (\( mg \text{ s}^{-1} \text{ m}^{-2} \)). The output variable, \( y \), is the odour intensity measured in odour units per second per meter squared (OU s\(^{-1} \) m\(^{-2} \)), and \( N \) is the number of samples. These statistical models are not capable of incorporating non-numeric or subjective human expert knowledge. The linear multiple regression model (LMR1) derived using two-component analysis (ammonia and hydrogen sulphide) of the odour is:

\[ y = 5.76 + 0.441x_1 + 0.0177x_2 \]  

2.4. Neural network models

Neural networks using supervised learning were chosen to model the pork farm odour for several important reasons. First, neural networks are a proven modelling technique that has been used to successfully model other complex biological systems where conventional statistical approaches have failed. Pork farm odour is a complex system that a neural network is able to model with little expert knowledge of the system. Second, the available dataset is very complete, containing all the input and output data necessary to utilize supervised learning. Reinforcement learning or unsupervised learning would have been considered had the available dataset been incomplete or too small to provide an adequate training set [3].

The two neural network models have been designed with three layers (input layer, hidden layer and output layer). The first neural network model (NN1) takes a multiple-component multiple-factor analysis approach to the odour and has two input neurons for the air-phase concentration of ammonia and hydrogen sulphide. The second neural network model (NN2) takes a multiple-component multiple-factor analysis approach to the odour and has five input neurons. The input neurons are the air-concentrations of ammonia and hydrogen sulphide, the time of day, month of the year, and the outdoor temperature. All the neural networks developed for this application have a single output neuron for the odour intensity. The network architecture for NN1 is shown in Fig. 1 and the network architecture for NN2 is in Fig. 2. The network architecture for NN1 consists of two input neurons, 16 hidden neurons and one output neuron. The NN2 has five input neurons, 20 hidden neurons and 1 output neuron. Many other factors, such as wind speed and relative humidity, which contribute to the pork farm odour, could also be used in the neural network as inputs if that information were available.

The inputs to the neural network and target during training have been normalized to the interval [0 1]. The largest target value in a dataset is the linear scalar function for the network output.

A three-layer neural network requires an activation function for the hidden layer and another for the output layer. The hidden layer needs a non-linear activation function and the output layer can use a linear or non-linear activation function. The non-linear activation function for the hidden layer is

\[ y = \tan(x) \]  

The linear activation function,

\[ y = x \]  

was used for the output layer of all neural network models developed in this investigation.
The feed-forward neural network models have been trained using a fast back-propagation (fast BP) learning algorithm [10]. The fast BP algorithm error function is defined as:

\[
E = \frac{1}{2} \sum_{k=1}^{N_o} (e_k)^2 = \frac{1}{2} \sum_{k=1}^{N_o} (y_t^k - y_k)^2
\]

(13)

where the predicted output is \( y \), the target output is \( y_t \), \( k \) is the index number of the output, and \( N_o \) is the total number of output neurons. The fast BP learning algorithm modifies the weights between neurons by the following formulae [10]:

\[
v_{kj} = v_{kj} + \alpha e_k^h h_j, \quad j = 0, 1, \ldots, N_h; \\
v_{kj} = 1, 2, \ldots, N_o
\]

(14)

\[
w_{ji} = w_{ji} + \alpha e_j^h x_i, \quad i = 0, 1, \ldots, N_i; \\
\]

\[
j = 1, 2, \ldots, N_h
\]

(15)

where \( x_i \) is the \( i \)th network input, \( h_j \) is the \( j \)th output of the hidden neuron, \( w_{ji} \) is the connection weight from the \( i \)th input neuron to the \( j \)th hidden neuron, \( v_{kj} \) is the connection weight from the \( j \)th hidden neuron to the \( k \)th output neuron, \( N_i \) is the total number of input neurons, and \( N_h \) is the total number of hidden neurons. Variable \( e_k^0 \) is the error of the \( k \)th output neuron; and \( e_j^h \) is the error of the \( j \)th hidden neuron. These parameters are calculated by [10]:

\[
e_k^0 = \lambda e_k + (1 - \lambda) \tanh (\beta e_k), \quad k = 1, 2, \ldots, N_o
\]

(16)
\[ \delta_j^h = (1 - \gamma_j^2) \sum_{k=1}^{N_h} \delta_k^{h+1} \delta_{kj}, \quad j = 1, 2, \ldots, N_h \]  
(17)

\[ \lambda = e^{-\mu/E^2} \]  
(18)

where \( \lambda \) and \( \mu \) are modification factors to determine the learning behaviour.

The trial number of hidden neurons for the neural network models was calculated using the equation below, which was used to determine the number of hidden neurons \( N_h \) for each network through an iterative process [20]:

\[ N_m = 0.5(N_i + N_o) + \sqrt{N_t} \]  
(19)

where \( N_i \) is the number of data points, and \( N_m \) is the trial number of hidden neurons for the network.

3. Results and discussion

It is crucial that a neural network model has the appropriate number of hidden neurons to achieve the best possible performance; this is discussed in the following section. Next, the performance results for models LMR1 and NN1, which only considered multiple-component analysis, are presented and compared against each other. Finally, the performance results for NN2, which used multiple-component multiple-factor analysis is presented and then contrasted with the results of NN1.

3.1. Number of hidden neurons

A three-layer neural network is theoretically able to learn any non-linear relationship at a desired level of precision with sufficient number of hidden neurons in the network architecture and through sufficient learning. An inadequate number of hidden neurons will result in under-fitting, or the inability to achieve the desired level of precision from the network [20]. Too many hidden neurons, known as over-fitting, will cause the dataset to be memorised and poor generalisation to be achieved [20]. Thus, it is very important that the appropriate number of hidden neurons is used in the neural network models. The trial number of hidden neurons \( N_m \) was calculated with Eq. (19) for both neural networks and used as a starting point. The number of hidden neurons was varied around \( N_m \) on a range of negative ten to positive ten hidden neurons with an increment of one. The number of hidden neurons \( N_h \) for each neural network model was determined through the comparison of \( E_a \), \( \sigma \), and \( R^2 \) for each neural network model in a trial and error approach. The appropriate number of hidden neurons for NN1 and NN2 was determined to be 16 neurons and 20 neurons, respectively.

3.2. Multiple component models

The models LMR1 and NN1 were both constructed and tested using the same training and testing sets. Both models considered the air-phase concentrations for ammonia and hydrogen sulphide as inputs or a multiple-component analysis approach. The neural network model NN1 achieved better performance than the statistical model LMR1 in every measure considered by this study, the results are presented in Table 1 and a graphical comparison of results are shown in Figs. 3 and 4. As shown in Table 1, the statistical model LMR1 realized an \( E_a \) of 64.9%, \( \sigma \) of 67.7%, and \( R^2 \) of 0.48. Whereas the multiple-component neural network model NN1 achieved an \( E_a \) of 31.2%, \( \sigma \) of 20.5%, and \( R^2 \) of 0.83. The mean absolute error for NN1 is less than half that of the error achieved by LMR1, indicating that the linear multiple regression odour model, LMR1, is less accurate than the corresponding multiple-component neural network. The much lower standard deviation attained by NN1 compared to LMR1 demonstrates that the neural network model is much more precise than the statistical model. The predicted odour intensity values generated by neural network model NN1 were highly correlated to the actual odour intensity values, \( R^2 = 0.83 \). While the odour intensities produced by the statistical model did not correlate as highly to the

<table>
<thead>
<tr>
<th>Model</th>
<th>( E_a ) (%)</th>
<th>( \sigma ) (%)</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMR1</td>
<td>64.9</td>
<td>67.7</td>
<td>0.48</td>
</tr>
<tr>
<td>NN1</td>
<td>31.2</td>
<td>20.5</td>
<td>0.83</td>
</tr>
<tr>
<td>NN2</td>
<td>19.3</td>
<td>12.7</td>
<td>0.81</td>
</tr>
</tbody>
</table>

\( E_a \): mean absolute error; \( \sigma \): standard deviation; \( R^2 \): coefficient of determination; LMR1: linear multiple regression model; NN1: multiple component neural network model; NN2: multiple-component multiple-factor neural network model.
actual odour intensities, $R^2 = 0.48$. Multiple-component analysis of the pork farm odour using neural network models improves the performance of the odour intensity predictions compared with the statistical model. The relationship that exists between the selected odour components and the odour intensity is a complex relationship that is non-linear. The neural network is able to automatically learn this non-linear relationship, whereas the researcher must select the statistical distribution to apply, which could be very difficult. Therefore, it is not unexpected that the neural network model has realized better performance than the corresponding statistical model.

The introduction of other significant odour components as inputs, such as indole and skatole, to these models would further increase the performance of both the statistical model and neural network model. It is highly probable that with the addition of other components, the neural network model would still surpass the performance of the corresponding statistical model due to the inherent learning ability of the neural network.

Other researchers have also moved towards multiple-component analysis of the pork farm odour, such as Zahn et al. [19] who have developed a nine component statistical model of the odour. To the best of our knowledge, no other researchers have attempted to model the pork farm odour using neural networks.

It has been demonstrated that the overall performance of multiple-component models for pork farm odours can be improved through the use of neural networks over statistical models. Based on our investigation and literature survey, we hypothesize that multiple-component pork farm odour models could be improved upon through the use of multiple-component multiple-factor analysis.

3.3. Multiple-component multiple-factor analysis

Previous studies on the modelling of pork farm odour have concentrated on the modelling of the odour components and have disregarded other potential contributing factors, such as the time of day and season, due to the limitation of statistical modelling techniques. The advent of soft computing techniques, such as neural networks, allows the incorporation of these other potential contributing factors or multiple-component multiple-factor analysis of the pork farm odour. The neural network model NN2 considers the time of day, month of year and outside temperature in addition to the components ammonia and hydrogen sulphide. The multiple-component multiple-factor neural network model NN2 delivered remarkable performance gains over the multiple-component neural network model NN1. There is a performance gain in $E_a$ of 31.2% for neural network model NN1 to 19.3% for neural network model NN2, a difference of 11.9%. Precision has also increased through the incorporation of the time of day, month of the year, and outside temperature. The standard deviation decreased modestly from 20.7% to 12.7% for models NN1 and NN2, respectively. The coefficient of determination remained approximately the same for both neural network models; $R^2$ is 0.81 for NN2 and 0.83 for NN1. The results for models NN1 and NN2 are presented in Table 1 and compared in Figs. 3 and 4. These performance gains indicate that factors other than the odourants are relevant to the analysis of the pork farm odour. Our literature survey revealed numerous factors that could be considered during
modelling, including relative humidity, dust levels, wind velocity, season, and ambient temperature [21].

It is hypothesized that through the inclusion of more components and contributing factors in the model, there will be greater performance achieved in terms of accuracy and precision for predicted odour intensities. A refinement of these factors to specific physical measures such as relative humidity from general aggregate factors, such as the time of day, will possibly improve the performance of model predictions. Pork farm odour is a very complex system with over 200 different components and a variety of factors that contribute to its generation. A more comprehensive neural network model would allow the identification of significant odour components and major factors that contribute to the odour generation using the “learning with forgetting” neural network method [6]. A comprehensive multiple-component multiple-factor model of pork farm odour could act as the basis for the development of a reliable pork farm odour measurement tool. An enhanced electronic nose system that measures air-phase concentrations of selected odourants and other physical properties, such as wind speed and relative humidity, could be integrated with the comprehensive model to provide real-time odour intensity measurement.

4. Conclusions

A linear multiple regression model and neural network model were developed based on multiple-component analysis of the odour. These two models used the air-phase concentration of ammonia and hydrogen sulphide as model inputs. The multiple-component neural network model provided better performance than the corresponding linear multiple regression model. This demonstrated that multiple-component pork farm odour models benefit from the use of contemporary intelligent modelling techniques, specifically neural networks. Despite the superior results of the neural network model, the addition of other significant odour components would probably increase the accuracy and precision of odour intensity predictions.

Through literature survey, it was anticipated that a multiple-component multiple-factor analysis approach to the pork farm odour would yield more accurate odour prediction models. As a result, a second neural network model was developed that considered several factors that contribute to the generation of the odour (time of day, month of the year, and outside temperature) in addition to the air-phase concentration of the two components. This multiple-component multiple-factor neural network model was compared against the multiple-component neural network model, and significant performance gains were realized in terms of the mean absolute error and the standard deviation. The consideration of odour components and generation factors in the modelling process, combined with intelligent modelling techniques will provide performance advances over conventional multiple-component statistical odour models. A comprehensive odour model could be used to identify significant odour components and factors of the odour generation process.

Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council (NSERC), Ontario Pork, and Ontario Ministry of Agriculture and Food (OMAF), Canada. The Department of Biosystems and Agricultural Engineering at the University of Minnesota, USA graciously provided the data used in this work.

References