University of Alberta

Risk Assessment of *Cryptosporidium parvum* Using Neural Networks

by

Kevin Robert Janes ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Electrical and Computer Engineering

Edmonton, Alberta
Fall 2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-33263-4
Our file Notre référence
ISBN: 978-0-494-33263-4

Abstract

Cryptosporidium parvum is a waterborne pathogen that has caused a significant number of outbreaks worldwide. A risk assessment of *Cryptosporidium parvum* exposure through drinking water requires consideration of oocyst concentrations at source waters, the effectiveness of drinking water treatment, tap water consumption, and the dose response relationship. In this thesis neural network models were developed to model tap water consumption and the disinfection of *Cryptosporidium parvum* using chlorine dioxide and ozone. These models were used for exposure assessment to determine daily doses of oocysts from tap water consumption. A dose response neural network model was developed for several strains of *Cryptosporidium parvum*. A risk characterization that considered a variety of exposure scenarios was completed using the tap water consumption, disinfection and dose response neural network models. The risk characterization produced point estimates for the daily probability of infection assuming exposure to *Cryptosporidium parvum*.

Acknowledgements

I would like to thank my supervisor Dr. Petr Musilek for his support and guidance. For reviewing this thesis and providing comment I would like to thank Joelle Hatton, Dr. Marek Reformat and Dr. Stanislav Karapetrovic.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 Problem Statement and Thesis Objectives

Waterborne *Cryptosporidium parvum* (*C. parvum*) is a microbial pathogen that affects the majority of the world's population through consumption of contaminated drinking water. Very compelling evidence that *Cryptosporidium* exposure via drinking water is occurring and causing infection is the number of documented waterborne *Cryptosporidium* outbreaks. Consider the 1993 outbreak in Milwaukee, Wisconsin that infected 403 000 people and killed over 50 (Rose *et al.*, 2002). Also, *C. parvum* has been shown to: have a low median infectious dose for humans, be resistant to various treatment processes and have the ability to survive under very harsh conditions for extended periods. These factors make *Cryptosporidium parvum* a considerable health risk.

The risk of infection to a person from *C. parvum* can be determined given point estimates of drinking water consumption, oocyst concentration, and dose response parameters. Determining the risk of infection requires the analysis and integration of many complex systems and application of many assumptions and models. Water consumption in risk assessments of waterborne pathogens is often assumed to be the same for all regions, ages, and genders even though it can vary considerably. The concentration of oocysts in treated drinking water is a function of land-based activities in a given watershed, climate and drinking water treatment processes. Individual drinking water treatment processes, such as disinfection, are very complex non-linear processes that are typically modeled using mathematical models that require deep process knowledge in order to develop. Also, mathematical models do not allow easy incorporation of additional input variables as new process knowledge becomes available.

There are developed quantitative risk assessment models for waterborne *C. parvum* that predict the probability of infection given a specific exposure to the pathogen through consumption. However, these models rely on conventional modeling techniques such as statistical and mathematical models that are not able to perform the best generalization of the dose response and exposure assessment relationships for a given dataset. For example, current dose response models for *C. parvum* are statistical and require the correct selection of a statistical distribution. The first major objective of the research described in this thesis was to overcome the limitations of current dose response models through the development of an intelligent dose response model for *C. parvum* based on neural networks. The intelligent dose response model produces a better generalization of the dose response relationship automatically, incorporates multiple strains of the *C. parvum* pathogen into the same model as input parameters and allows for quantitative risk assessments of immuno-deficient populations, and populations with previous *C. parvum* infection and the resulting immunity. The intelligent dose response model is compared against several exponential dose response models.

The second objective of the research is to use neural network modeling techniques for the exposure assessment of waterborne *C. parvum* while considering several disinfection water treatment processes. Three separate factors are considered in the exposure assessment:

1. The presence of the *C. parvum* oocysts in raw water sources through various contamination sources and factors;
2. The effect of drinking water treatment disinfection processes on *C. parvum* oocysts, and;
3. Consumption rates of drinking tap water.

The third objective of the research is to perform a quantitative risk characterization of waterborne *C. parvum* using the developed neural network dose response and exposure assessment models to determine point estimates for the daily probability of infection from *C. parvum*. The risk characterization considers the following simulation conditions: various oocyst concentrations, varying disinfection process effectiveness, several drinking water consumption rates, and several different *C. parvum* strains.

1.2 Cryptosporidium

Cryptosporidium is a protozoan pathogen that develops into an oocyst (multiple parasites contained in a double walled outer shell) before it is excreted in the feces of the infected host (Fayer *et al.*, 2000). This oocyst stage is very important for the distribution, endurance, and transmission of the *Cryptosporidium* parasite. *Cryptosporidium* oocysts are able to remain active for extended periods of time (several months), even under harsh conditions ranging from –20 °C to +20 °C (Fayer *et al.*, 2000). However, very extreme cold (-70 °C) and extreme heat (+70 °C) applied to an oocyst for a short period, approximately 10 seconds, has been shown to kill oocysts. *Cryptosporidium* has been recognized as causing human disease since 1976 and has been generally recognized as a waterborne pathogen since 1984 (Hrudey and Hrudey, 2004).

The transmission pathway for *Cryptosporidium* starts when oocysts are ingested and possibly infecting the host. Subsequently, the infected host produces and excretes many oocysts to the environment and another host might ingest those oocysts. This is known as fecal-oral transmission. There are 10 different species of *Cryptosporidium* that have been identified, but *Cryptosporidium parvum* is the species considered to be the cause for cryptosporidiosis in humans (Fayer *et al.*, 2000; Rose *et al.*, 2002). *C. parvum* has been known to infect cattle, goats, swine, and mice in addition to humans (Martins and Guerrant, 1995). Once *C. parvum* has been ingested, the pathogen's life cycle is completed within the gastrointestinal tract of the host where it multiplies and many of the symptoms of cryptosporidiosis may appear. The symptoms include diarrhea, cramps, nausea and abdominal pains (Messner *et al.*, 2001). There is currently no cure or vaccination for cryptosporidiosis, there are only re-hydration treatments available for people that lose large amounts of body fluids while ill with cryptosporidiosis. In severe cases, the rapid loss of significant body fluids from diarrhea due to cryptosporidiosis can result in death. Death from cryptosporidiosis is more likely to occur in sensitive populations such as the elderly, the immuno-compromised, and the malnourished (Teunis *et al.*, 2002). For those with HIV, new developments in drug therapies have helped lower the risk of developing chronic cryptosporidiosis by protecting the immune system (Hrudey and Hrudey, 2004). A person who is exposed to *C. parvum* oocysts will not always develop an infection, however if an infection does occur they will produce and excrete oocysts (Joseph *et al.*, 2005).

There are a variety of methods for transmitting *Cryptosporidium* oocysts such as person-to-person, animal to person, foodborne transmission, and waterborne transmission. Waterborne transmission (fecal contaminated water) is considered to be the primary source of transmission for *Cryptosporidium* (Donnelly and Stentiford, 1997). The contaminated fecal matter of humans and animals has many opportunities to enter the water system. Generally, groundwater aquifers are not easily contaminated with *Cryptosporidium*, but surface water sources like rivers, streams and lakes are easily contaminated. Human wastewater does receive treatment such as screening, settling and disinfection in many developed nations. However, the oocysts are resistant to many of these processes, specifically chlorine disinfection, and most developing nations have minimal processing of human wastewater. In developed nations, with the advent of factory farming and modern agricultural practices, the untreated fecal waste of many types of livestock is concentrated, liquefied and spread on to farming lands as fertilizer and the excess runs off into surrounding surface waters. Finally, wildlife does produce fecal matter that contributes to the water contamination problem, but wildlife contribution is low relative to other sources. Every human must drink water, it is essential to life; thus, it is not surprising that waterborne transmission is the major transmission pathway for *Cryptosporidium* in humans. Waterborne pathogens like *Cryptosporidium* use many transmission pathways in combination, which increases their posed threat to the human population (Joseph *et al.*, 2005).

1.3 Prevalence in the Environment and Exposure

In order to evaluate the risk associated with waterborne *Cryptosporidium*, it is necessary to consider the infectivity of the pathogen, potential adverse health effects, transmission pathways, and also the potential exposure of the public to the parasite. There are a variety of factors that influence potential *Cryptosporidium* contaminated drinking water exposures, including:

- The concentration of oocysts in drinking water sources;
- Prevailing weather conditions;
- The virulence and robustness of the oocysts;
- Oocyst removal or deactivation by water treatment processes, and;
- The amount of cold water consumed on a daily basis.

There are two types of drinking water sources, surface water and groundwater. It is known that surface water sources are more vulnerable to fecal contamination than groundwater, but groundwater sources have also been contaminated with *C. parvum* oocysts at relatively low concentrations in the past (Rose *et al.*, 2002).

The excretion rate of oocysts from infected animals and people surrounding a water source is very important in determining the loading for the water source. The land use for areas surrounding the water source is an essential factor contributing to the concentration of oocysts in water sources. It has been shown that source waters with extensive agriculture surrounding the water body or a wastewater treatment outlet nearby are likely to have 10 to 100 times the concentration of oocysts than water sources that do not (Rose *et al.*, 2002). This is of particular concern to areas with a combination of high population densities, extensive agriculture and heavy reliance on surface sources for drinking water, such as New Zealand and the Netherlands (Duncanson *et al.*, 2000; Teunis *et al.*, 1997). This results in extraordinarily heavy loading of the water source due to livestock fecal and fertilizer runoff, and wastewater discharge. It is reported by a variety of nations, including Canada, USA, UK, Spain, and New Zealand, that the ranges of average concentration of *C. parvum* oocysts in wastewater is 3000 to 4000 oocysts per litre of water, surface water is 12 to 250 *C. parvum* oocysts per litre of water, and drinking water is 0.5 to 6 oocysts per litre of water (Makri *et al.*, 2004; Rose *et al.*, 2002; Teunis *et al.*, 1997). This illustrates that large numbers of oocysts are being released via wastewater discharge and that surface waters have relatively high concentrations of oocysts. Also, treated drinking water may contain roughly enough oocysts per litre to infect a person depending on the strain of *C. parvum*, the person's individual susceptibility, and the amount of water ingested per day. The reported amount of drinking water ingested per day in North America ranges from 1 to 3 litres of water depending on age, lifestyle and other factors (Makri *et al.*, 2004; Teunis *et al.*, 1997). Teunis *et al.* (1997) calculate that the median individual probability of infection for the *C.*

parvum Iowa strain per annum is approximately 1 in a population of 10 000 for healthy individuals who drink 2 litres of water daily from a surface water source.

Regarding *C. parvum* outbreaks, Canada had four documented waterborne *Cryptosporidium* outbreaks between 1993 and 1996 that affected almost 32 000 people. The alleged source of contamination in most cases was wastewater discharge and liquid manure fertilizer runoff (Rose *et al.*, 2002). The average hospitalization rate from *Cryptosporidium* in North America and the United Kingdom is reported to be approximately 13% among those who become infected (Rose *et al.*, 2002). That would mean over 4000 hospitalizations in Canada for documented *C. parvum* outbreaks between 1993 and 1996. The United States had 10 documented waterborne *Cryptosporidium* outbreaks between 1984 and 1996, while the United Kingdom had 13 documented outbreaks between 1986 and 1996 (Rose *et al.*, 2002).

1.4 Risk Assessment

Risk analysis encompasses risk assessment, risk management and risk communication. Risk assessment can be defined as:

> "A systematic process for qualitative or quantitative characterization of adverse effects (risks) associated with hazardous substances, processes, actions, and/or events" (Gibson *et al.*, 1998).

Risk management is the process of developing a risk treatment strategy that mitigates the risk identified during the risk assessment and takes into consideration financial, political, and other constraints. Risk communication is the communication of identified risks and the risk treatment strategy, if applicable, to stakeholders. Quantifiable risk assessment has traditionally focused on human exposure to chemicals, in particular carcinogenic substances. In the last 25 years quantitative microbial risk assessment (QMRA) for waterborne and foodborne pathogens has emerged (Gibson *et al.*, 1998).

Quantitative microbial risk assessment is considered an important tool for understanding and managing the risk from waterborne pathogens, such as *C. parvum* (WHO, 2004). Quantitative microbial risk assessment for waterborne pathogens involves four major steps: hazard identification, exposure assessment, dose response assessment, and risk characterization. The risk assessment process is presented in Figure 1.1. Hazard identification involves discovering all possible drinking-water hazards that may expose human consumers to waterborne pathogens (WHO, 2004).

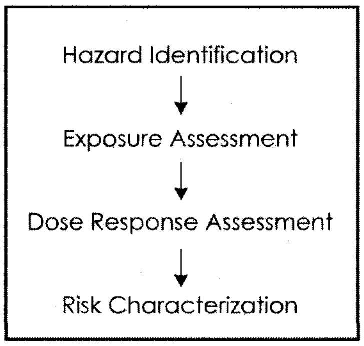

Figure 1.1: Risk assessment process

Using tap water as an example, exposure assessment for waterborne pathogens is the assessment of how many pathogenic organisms the exposed individual will have come in contact with through ingestion of contaminated tap water. Dose response assessment characterizes the incidence of an adverse health effect in the population given a certain level of exposure to a pathogen. Risk characterization involves integrating the information generated by the exposure and dose response assessments to arrive at an estimation and description of the risk (Thomas and Hrudey, 1997).

1.4.1 Hazard Identification

For waterborne pathogens, hazard identification involves identifying a particular pathogen's potential to cause significant adverse health effects within a population. Information collected during the hazard identification step includes information about the pathogen, such as survival and growth conditions and potential transmission pathways for example. Epidemiological studies, outbreak data, and surveillance data provide excellent information sources for hazard identification. Epidemiological studies are either observational or experimental and involve human subjects. Observational studies involve investigators measuring the natural evolution of a disease without intervention, while experimental studies involve direct intervention by the investigators through the exposure of a pathogen to a subject or the progression of the disease caused by a pathogen (Beaglehole *et al.*, 1993). Surveillance data is derived from the continuous monitoring of disease occurrence within the population. In general hazard identification is not considered a major component of QMRA, while exposure assessment, dose response assessment, and risk characterization are significant steps.

A critical component of hazard identification involves deciding what primary consequence will be measured by the risk assessment: infection, illness, or death. Generally infection within the population has been regarded as the primary consequence that should be protected against (WHO, 2001). For this thesis, infection was the only consequence considered for the risk assessment. Hazard identification for *C. parvum* was previously discussed in sections 1.2 and 1.3.

1.4.2 Exposure Assessment

Exposure assessment for this research describes a particular transmission pathway through which a waterborne pathogen goes from a raw water source to a drinking water consumer. The purpose of exposure assessment is to determine the particular mechanism for exposure and to estimate the level of exposure. The mechanism of exposure considered in this thesis for *C. parvum* is limited to drinking water consumption. When estimating the level of exposure to *C. parvum* through drinking water it is necessary to estimate the consumption of drinking water per day and the oocyst concentration at the point of consumption. When estimating the oocyst concentration one must consider the introduction and distribution of the pathogen in a raw water source, water treatment, and finally the distribution of drinking water to a consumer. Since *C. parvum* oocysts are only able to reproduce within a host, it is assumed for this thesis that the drinking water discharged from a drinking water treatment plant is reasonable for estimating the oocyst concentrations in the drinking water consumed. A study by Haas and Rose (1996) demonstrated that *Cryptosporidium* oocysts are uniformly distributed within in water sample. It is assumed for this thesis that *Cryptosporidium* oocysts are uniformly distributed within a water supply.

1.4.3 Dose Response Assessment

Dose response assessment uses the exposure information produced by the exposure assessment to estimate the risk of infection and possibly illness for the exposed population. Dose response assessment is dependent on the availability of data to produce a dose response relationship that translates a specific level of exposure into a health response within the affected population. Data from human dose response studies are generally the basis for dose response relationships, but data from animal studies may be used in the absence of available human information. The use of animal dose response data must be used with caution due to differing metabolisms and the resulting need to extrapolate results from the test species to humans.

1.4.4 Risk Characterization

In QMRA, risk characterization integrates the information generated from the exposure and dose response assessments to calculate an overall estimation of the risk. A point estimate of risk can be determined through a combination of point estimates for exposure (the number of pathogens consumed in cold drinking water) and dose response relationship parameters (WHO, 2001). Another approach is probabilistic risk assessment that considers the entire distribution of the exposure and dose response relationships instead of single point estimates, and then estimates the uncertainty and variability within the relationships and resulting risk estimate. The point estimate approach was used in the research described in this thesis.

1.5 Thesis Organization

The second chapter of this thesis provides a detailed description of the types of neural networks used for modeling and the methods used to evaluate and analyze the developed models. Chapter three reviews the concentration of *C. parvum* oocysts in source waters and outlines the concentration assumptions made in this thesis for the risk characterization. The fourth chapter of the thesis details the design and benchmarking of two neural network disinfection models (ozone and chlorine dioxide) against the temperature corrected Chick-Watson model. The fifth chapter presents two tap water consumption neural network models that are based on Canadian tap water consumption. Chapters three, four, and five represent the components needed for exposure assessment. A *C. parvum* dose response neural network model for three different strains is presented in chapter six and compared against the traditional exponential dose response model. Chapter seven contains the assumptions and results of a risk characterization for *C. parvum* using the developed neural network models explained in previous chapters. The eighth chapter outlines some of the limitations of the completed risk assessment, including parameter uncertainty, model uncertainty, variability and the correctness of assumptions and design decisions. Finally, the ninth chapter provides general conclusions and extensions for further work.

2.0 METHODS BACKGROUND

2.1 Neural Networks

The neural network is a computing paradigm that is modeled after the structures of the brain, where a number of interconnected processing units (neurons) collaborate to generate an output. Neural networks can be used to simulate and analyze complex systems that cannot be easily modeled by statistical or mathematical models. There are many learning algorithms available for training a neural network but the objective of all learning algorithms is to optimize the connection weights that interconnect the neurons. The method of optimization varies with each learning algorithm. The learning algorithms used in this research include backpropagation and structural learning with forgetting.

2.1.1 Overfitting and Underfitting[1]

The modeling capacity of a neural network is dependent on its ability to capture the most important features from data during training. The ability of the neural network to generalize the underlying relationship is crucial and can be affected by the overfitting and underfitting problems. Underfitting occurs when a neural network does not have a sufficient number of hidden units and is unable to detect the correct relationship from the dataset. Conversely overfitting occurs when the

[1] A version of this section has been accepted for publication.
Janes, K.R., and Musilek, P. In Press. Neural network models of *Cryptosporidium parvum* inactivation by chlorine dioxide and ozone. Journal of Environmental Engineering and Science.

network has many more hidden units than what is needed to detect the relationship in the dataset and simply memorizes the training set and its associated noise. The overfitting problem is particularly relevant for small datasets, such as those used in this research. There are a number of different methods available for avoiding both underfitting and overfitting. Three popular methods are early stopping, jittering, and weight decay. Early stopping requires splitting the dataset into three sets, a training set, test set and validation set. After an iteration of the training set, the neural network is assessed with the validation set. The network with the best results is used for testing with the test set. Jitter is simply training with noise added to the inputs on purpose. Finally, weight decay is a regularization method that avoids overfitting through the addition of a penalty term to the error function in a learning algorithm (Rognvaldsson, 1998). The objective of all these methods is similar, to improve the generalization achieved by the neural network. To avoid the overfitting and underfitting problems in the disinfection and water consumption neural network models, weight decay was employed through the selection of the "structural learning with forgetting" learning algorithm. Due to the small dataset available for developing the neural network dose response model it was necessary to qualitatively review the plotted dose reponse relationship for smoothness to avoid the overfitting problem. In a qualitative review, if the relation is not smooth then overfitting has occurred. Yang (2003) successfully used this technique in the development of a unified neural network dose response model for multiple foodborne pathogens. If overfitting occurred in this study it would mean that the developed neural network models might produce inaccurate predictions that were influenced by the noise in the training set detected by the overly complex neural network during training.

2.1.2 Backpropagation Learning[2]

The classic three-layer multilayer perceptron neural network with gradient descent based backpropagation training was used to develop the dose response neural network models presented in this thesis. The initial connection weights between layers of processing units and biases were randomized in the interval [-1, 1]. The activation function for the hidden layer had to be nonlinear; thus the unipolar sigmoidal function was used:

$$y = \frac{1}{1+e^{-x}}$$

The activation function for the output layer can be linear or nonlinear; a linear activation function was chosen for the output layer:

$$y = x$$

The inputs and outputs for training and inputs for testing the neural networks were normalized to the interval [0, 1]. The largest target output value in a dataset

[2] A version of this section has been published.
Janes, K.R., and Musilek, P. 2007. Modeling the disinfection of waterborne bacteria using neural networks. Environmental Engineering Science, **24**(4): 448-459.

was used as the linear scalar function for a model's output. The neural network output y is given by:

$$y_k = \sum_{j=0}^{n} w_{jk} h_j$$

where n is the number of hidden units, and w_{jk} is the connection weight between the j-th hidden unit and k-th output unit. The output of the j-th hidden unit, h_j, is given by:

$$h_j = \frac{1}{1 + \exp\left[-\left(\sum_{i=0}^{m} w_{ij} x_i\right)\right]}$$

where m is the number of input units, w_{ij} is the connection between the i-th input unit and the j-th hidden unit and x_i is the i-th input.

A gradient descent based algorithm was used for training by minimizing the following error function:

$$E = \frac{1}{2}\sum_{q=1}^{s}(d_q - y_q)^2$$

where E is the sum of squared errors, s is the number of data patterns, and d_q and y_q are the desired value and model output for the q-th data pattern. The connection weights are updated by the following equations:

$$w_{ij} = w_{ij} - \eta \left[\frac{\partial E}{\partial w_{ij}} \right]$$

$$w_{jk} = w_{jk} - \eta \left[\frac{\partial E}{\partial w_{jk}} \right]$$

where η is the learning rate. These backpropagation connection weight update rules are based on the gradient descent method, which makes use of the derivative to step in the direction that will yield the maximum decrease of the network error.

2.1.3 Structural Learning with Forgetting[3]

Neural networks have proven to be very successful for modeling many engineering, financial, and biological systems. Neural networks are capable of automatically determining the input-output relationships for these complex nonlinear systems. Despite many benefits, neural network based models do have disadvantages. First, neural networks require the selection of a network structure, which greatly influences the success or failure of the model. If a network has too many hidden neurons it will likely have poor generalization and an insufficient number of neurons will cause underfitting. Second, the inclusion of irrelevant variables in neural networks is frequent because no previous

[3] A version of this section has been accepted for publication.
Janes, K.R., and Musilek, P. In Press. Neural network models of *Cryptosporidium parvum* inactivation by chlorine dioxide and ozone. Journal of Environmental Engineering and Science.

knowledge is required for the neural network to determine input-output mappings for a system. The inclusion of irrelevant variables can greatly decrease model performance. Structural learning with forgetting (SLF) is a destructive learning method that addresses these issues by starting with a large fully connected network and then driving insignificant connection weights towards zero. The resulting lean structure indicates which hidden neurons can be eliminated and potentially which input and output neurons.

Structural learning with forgetting is a learning method based on the standard backpropagation learning algorithm that has been extended to include weight decay. Structural learning with forgetting consists of learning with forgetting, hidden units clarification and learning with selective forgetting. Learning with hidden units clarification attempts to promote localized representations within hidden units by forcing the units to be completely active, producing an output of one, or inactive, producing an output of zero. Hidden units clarification requires binary outputs, and was not included in any of the neural network models developed.

Learning with forgetting involves constant decay of connection weights through a penalty criterion, which eventually produces a skeletal structure. Ishikawa (1996) defines the criterion function in learning with forgetting to be:

$$J_f = J_{bp} + \varepsilon' \sum_{i,j} |w_{ij}|$$

where J_f is the total criterion for learning with forgetting, J_{bp} is the mean square error in back propagation learning, w_{ij} is a connection weight and the last term is the penalty criterion where $\varepsilon'$is its relative weight.

A disadvantage of learning with forgetting is that it produces larger mean squared errors than standard back propagation learning. This is addressed by selective learning with forgetting by only including weak connection weights below a specified threshold, β, in the total criterion (Ishikawa, 1996):

$$J_s = J_{bp} + \varepsilon' \sum_{|w_{ij}|<\beta} |w_{ij}|$$

where J_s is the total criterion for selective learning with forgetting.

The SLF algorithm modifies the connection weights, between layers of neurons according to (Ishikawa, 1996):

$$\Delta w_{ij} = -\eta \frac{\partial J_{bp}}{\partial w_{ij}} - \varepsilon \, \mathrm{sgn}(w_{ij})$$

where η is the learning rate, ε is a decay rate, Δw_{ij} is the weight change, and sgn is a sign function. The first term represents the normal weight change when only using back propagation learning, and the second term is the penalty for SLF.

In this study all neural network models developed with SLF:

- Used a nonlinear activation function for the hidden layer, and a linear activation function for the output layer;
- Used data for training and testing the neural networks that were normalized to the interval [0 1], and;
- Had the largest expected output value used as a linear scaling function for the model output.

2.2 Saliency Analysis[4]

Neural networks use a distributed representation of knowledge, where each connection weight and processing unit within the structure represents only a fraction of knowledge in the network (Dawson, 2004). Thus the network will produce acceptable outputs even when confronting incomplete or noisy input data due to this built-in fault tolerance. The intentional omission of input data to a trained network is known as saliency analysis, and provides a method for investigating the internal relationships between variables within the network. Saliency analysis establishes the relative significance of each input variable to the model by observing the effect that each omitted variable has on an error function (Abrahart *et al.*, 2001). By omitting the necessary number of inputs, single and multiple input saliency analysis can be performed. In this study only

[4] A version of this section has been published.
Janes, K.R., and Musilek, P. 2007. Modeling the disinfection of waterborne bacteria using neural networks. Environmental Engineering Science, **24**(4): 448-459.

individual inputs were omitted. Abrahart *et al.* (2001) suggest three methods for omitting inputs: setting appropriate weights to zero, setting the output values of the appropriate processing unit to zero, or zeroing the data inputs to an input unit. In very simple neural networks it is possible to perform a direct examination of connection weights within the network to discover the relative importance of each variable within the network. However, this method is extremely difficult to apply when the neural network becomes complex and, consequently, saliency analysis provides an alternative analysis method. In this study the omission of input variables was achieved by zeroing the data inputs for a particular input unit, and the mean absolute error was used as the error function. The effect on the coefficient of determination was also examined in this study; however, the effect on the mean absolute error was considered the primary indicator for the saliency analysis.

2.3 Evaluation Methods

The developed neural network models were evaluated based on analysis of prediction results generated from test sets. The data in the test sets were never used for training at any point in neural network model development. Baxter *et al.* (2002) recommend using absolute measures of error to allow easy comparison of errors to actual targets in model development. The performance for all the neural network models developed in this research has been evaluated using mean absolute error (MAE) and the coefficient of determination (R^2). The objective of each model is to minimize the mean absolute error and maximize the coefficient of determination. The MAE measures the mean of the absolute difference between predicted and observed values (Zhang *et al.*, 2002). The coefficient of determination was used to ascertain the correlation between observed values in the test set and model predictions. The coefficient of determination represents the percent of variation between model predictions and actual observations accounted for by the model. The coefficient of determination ranges from zero to one, where a value of zero indicates the model has no predictive capability and one means the model has perfect predictive capability.

3.0 *CRYPTOSPORIDIUM PARVUM* CONCENTRATIONS

The concentration of oocysts in raw drinking water sources has a direct impact on the concentration of oocysts in the final drinking water. *C. parvum* oocysts can be found in most surface water sources, and in some ground water sources in very low concentrations (Rose *et al.*, 2002). There are many variables that contribute to the concentration of *C. parvum* oocysts in the environment, including watershed management, wastewater treatment and climatic conditions such as rainfall.

Management of watershed land use plays a significant factor in the concentration of oocysts. Heavy agricultural land use or wastewater discharge along a surface water body is known to increase the concentration of oocysts (Rose *et al.*, 2002). Many significant *C. parvum* outbreaks have been associated with increased levels of rainfall (Rose *et al.*, 2002). A study of several Washington state rivers demonstrated that oocyst concentrations were higher during periods with high rainfall runoff than dry periods (Atherholt *et al.*, 1998).

The prediction of oocyst concentrations at the intake of drinking water treatment plants would provide treatment plant operators with early warning of poor raw water conditions. Most studies have sought to find statistically significant associations between oocyst concentrations with climate and water quality parameters. However, the correlation of water quality parameters and the concentration of a pathogen can vary from site to site due to different environmental conditions (Atherholt *et al.*, 1998), thus making it difficult to develop a universal method for providing early warning to operators.

Brion *et al.* (2001) used neural networks to investigate the relationship between several water quality and quantity variables at the intake of a drinking water treatment plant in order to predict the peak concentration of *C. parvum* oocysts entering the treatment plant. Brion *et al.* (2001) found that turbidity was the least significant parameter for predicting concentrations, while the concentration of *Clostridium perfringens* was determined to be the most significant for the selected model input parameters. This study found that neural network models were able to predict peak concentrations of oocysts at the intake, but the development and application of the neural network models must be on a site-specific basis for predicting the concentration of oocysts in the environment. It was noted that to determine the appropriate combination of model inputs to permit application of a neural network model on multiple sites would require significantly more raw data from many sites with differing climatic conditions.

3.1 Concentration Assumptions

The raw data required to develop a neural network model capable of predicting oocysts concentrations on many sites was not available. For the simulation of a water treatment process in this study, a range of oocyst concentrations in surface water sources identified in the literature were used. Rose *et al.* (2002) summarized the findings of several water monitoring studies for *C. parvum* and found that the concentration of oocysts in surface water ranged from 12 to 250 oocysts per liter. Concentrations within this range were monitored in Australia, Germany, Israel, Malaysia, Netherlands, Spain, United Kingdom and the United States of America. Three concentrations of oocysts were assumed and used for simulations in this thesis: 12, 119, and 250 *C. parvum* oocysts per liter of water. This is intended to represent a low, medium and high loading of oocysts within raw surface water sources.

4.0 NEURAL NETWORK DISINFECTION MODEL

4.1 Background[5]

Waterborne disease caused by *C. parvum* is a significant problem for the populations of developed and developing nations. *C. parvum* oocysts are extremely difficult to inactivate due to the robust outer shell protecting the pathogens while in transit from host to host. There are a number of competing disinfectants that can be used with varying results to inactivate *C. parvum* oocysts. Two effective disinfectants for *C. parvum* are chlorine dioxide and ozone. The Chick-Watson (CW) model, a combination of theories from Chick (1908) and Watson (1908), is a popular disinfection model that has been used to characterize the inactivation relationships for both of these disinfectants (Li *et al.* 2001a; Li *et al.* 2001b; Li *et al.* 2001c). This mathematical model has achieved good prediction capabilities; however there are alternative computational modeling techniques such as artificial neural networks that can decrease prediction errors with available experimental data. Other unit processes in drinking water treatment such as coagulation have already been successfully modeled using neural networks (Baxter *et al.* 2002). The inactivation of a protozoan pathogen, *Giardia lamblia*, by chlorine has been modelled using neural networks by Haas (2004). Heck *et al.* (2001) used artificial neural networks to

[5] A version of this chapter has been accepted for publication.
Janes, K.R., and Musilek, P. In Press. Neural network models of *Cryptosporidium parvum* inactivation by chlorine dioxide and ozone. Journal of Environmental Engineering and Science.

model the inactivation of a parvovirus using ozone. Heck *et al.* (2001) used six input conditions in the developed model: alkalinity, initial virus concentration, sonication, organic carbon concentration, time, and ozone residual.

Janes and Musilek (2007) performed a detailed comparison of multi layer perceptron disinfection models for *Escherichia coli* and *Eberthella typhosa* trained using backpropagation (local optimization) and simulated annealing (global optimization). Both sets of models produced similar performance results and were found to be functionally equivalent. However, the models trained using simulated annealing required significantly more development time than the backpropagation trained models. Janes and Musilek (2007) concluded that deterministic derivative-based learning for disinfection models would likely perform well against other stochastic derivative-free learning methods such as genetic algorithms and require less development time. For the neural network disinfection models of *C. parvum* developed in this research, deterministic derivative-based learning algorithms were used.

The development of an appropriate network structure is very difficult due to the overfitting and underfitting problems, and identification of relevant model inputs. These issues can be addressed by structural learning with forgetting, a destructive learning method that starts with a large fully connected network and then drives insignificant connection weights towards zero. The resulting sparse structure indicates which hidden neurons can be eliminated and potentially which input and output neurons. The neural network disinfection models presented here have been trained using SLF.

Haas (2004) pondered how neural networks could consider the decay of the disinfectant residual. This is not a large consideration for a disinfectant like chlorine that has a relatively stable residual, whereas it is highly relevant for ozone, which is very reactive and dissipates quickly. A simple solution is to include both the initial residual and final residual concentrations for the disinfectant as input parameters to the neural network model. This approach has been taken in this study for both chlorine dioxide and ozone. The objectives of the work in this section are: to establish the performance of neural network models for disinfection of waterborne cryptosporidium relative to existing Chick-Watson models; demonstrate that the inclusion of the final residual concentration is an effective approach for considering residual decay; and rank the input parameters. A thorough literature review found that the inactivation relationships of waterborne *C. parvum* using chlorine dioxide and ozone have never been modeled using neural networks.

4.2 Materials and Methods

4.2.1 Datasets

It is very expensive and effort intensive to perform disinfection experiments, as a result both of the datasets used in this study are relatively small for training a neural network. Using small datasets means that the selection of the test set is much more influential to model outputs and consequently model performance evaluation. This issue can be mitigated by randomly selecting the data for each set while balancing the representation of data on the training set and test set to the best degree possible through statistical analysis of each dataset. This approach was taken in this study by randomly selecting and assigning data patterns to the training and testing sets until the arithmetic mean of the input and output variables were within 50% of each other for the two subsets. Another disadvantage of using a small dataset is that the neural network is more susceptible to overfitting, where the network simply memorizes the training set and its underlying noise (Silvert and Baptist, 2000). The risk of overfitting has been mitigated in this study through the selected learning algorithm SLF, which has been discussed in section 2.1.3.

It was not possible to test for the experimental error in the underlying datasets, as there were no deliberate experimental replicates available in the literature to test this. It is assumed that the experimental error is constant and consistent for the datasets used. The same research group at the University of Alberta produced both datasets. Both datasets used for model development were analyzed for missing data and outliers. Any data patterns within a dataset that were missing data were excluded from the dataset. However, this did not occur for either dataset. For outlier detection all data patterns that had input or output values of ±3 standard deviations from the mean of a model would have been removed, but this was not necessary for either dataset.

Li *et al.* (2001a) studied chlorine dioxide inactivation of *C. parvum* oocysts at pH levels of 6, 8, and 11 and a temperature range from 1°C to 37°C. The study by Li *et al.* (2001a) produced 61 data points, each containing the initial residual and final residual (concentration at the end of the contact time) in milligrams per liter, water temperature in degrees Celsius, contact time in minutes, the pH, and the observed inactivation ratio in log units. The initial residual ranged from 0.39 mg/L to 6.05 mg/L and the contact time ranged from 15 minutes to 240 minutes. For model development 48 data points were randomly selected for training and 13 for testing.

The dataset for ozone inactivation of *C. parvum* was compiled from two separate studies, Gyurek *et al.* (1999) and Li *et al.* (2001c). The data from both studies were generated under similar experimental conditions by the same research team in Alberta, Canada. A total of 62 data points were available for model development and testing, 13 were randomly reserved for testing and 49 for network training. Each data point contains the initial and final residuals, temperature, contact time, pH and inactivation ratio. These inactivation studies used a temperature range of 1°C to 37°C and pH levels of 6, 7, and 8. The contact time varied from 2 to 30 minutes, and the initial residual ranged from 0.3 mg/L to 2.7 mg/L.

The complete datasets used in this study are available in the following papers: Li *et al.* (2001a) for chlorine disinfection, and Li *et al.* (2001c) and Gyurek *et al.* (1999) for ozone disinfection.

4.2.2 Chick Watson Models

Assuming a constant disinfectant residual the classic Chick-Watson model is (Li *et al.* 2001b):

$$\log \frac{N}{N_o} = -kC^n t$$

Where t is the contact time in minutes, C is the constant residual, k and n are empirical constants found through experiment, N_o is the initial microorganism concentration, and N is the concentration of microorganisms at time t. However, this simple model requires a separate set of empirical constants for each pH and temperature combination and does not consider residual decay. A temperature corrected Chick-Watson model, which assumes n is one, has been used for comparison with the developed neural network models:

$$\log\frac{N}{N_o} = -k_{22}\theta^{T-22}\left(\frac{C_0 - C_f}{2}\right)t$$

where C_o is the initial disinfectant residual, C_f is the final disinfectant residual, k_{22} is the inactivation rate constant at 22°C, T is the temperature in °C and θ is the temperature coefficient. Li *et al.* (2001b) used this modified Chick-Watson equation to model the inactivation of *C. parvum* using ozone and chlorine. The empirical constants used in this study for the chlorine dioxide inactivation of *C. parvum* are k_{22} = 0.018 and θ=1.085. Li *et al.* (2001a) developed these constants based on the 61 data points used for training the neural network models in this study and as a result the applicable temperature range is from 1°C to 37°C and the pH range is 6 to 11 even though pH is not explicitly considered as a model input. The empirical constants used in this study for the ozone disinfection of *C. parvum* are k_{22} = 0.39 and θ=1.104, and were developed by Li *et al.* (2001c) based on the 62 data points used for model training and testing. The ozone model is considered valid for temperatures of 1 to 37°C and pH 6 to 8. In both studies by Li *et al.* (2001a; 2001c) all of the available data was used to develop the temperature corrected Chick-Watson models, none of the data was reserved for model testing. In this study a portion of the datasets were reserved for model testing purposes. The temperature corrected Chick-Watson model for inactivation of *C. parvum* using chlorine dioxide will be referred to as CW1 and the ozone disinfection model will be referred to as CW2.

4.2.3 Neural Network Models

The ability of the neural network to generalize the underlying relationship is crucial and can be affected by the overfitting and underfitting problem. The approach taken in the development of the disinfection models to avoid the overfitting and underfitting problems was weight decay through the selection of SLF as the learning algorithm.

All neural network disinfection models in this study consist of an input layer, hidden layer, and output layer, where there are five input units and one output unit (Figure 4.1). The input units are initial residual, final residual, contact time, pH and water temperature; the output unit is the inactivation rate in log units. The chlorine dioxide and ozone disinfection neural network models are referred to as CDNN and ONN respectively. Both CDNN and ONN commenced training with 10 hidden units and the SLF algorithm found the optimal number of hidden units to be four and three respectively. None of the input parameters became redundant during training, each maintaining several significant connection weights after training.

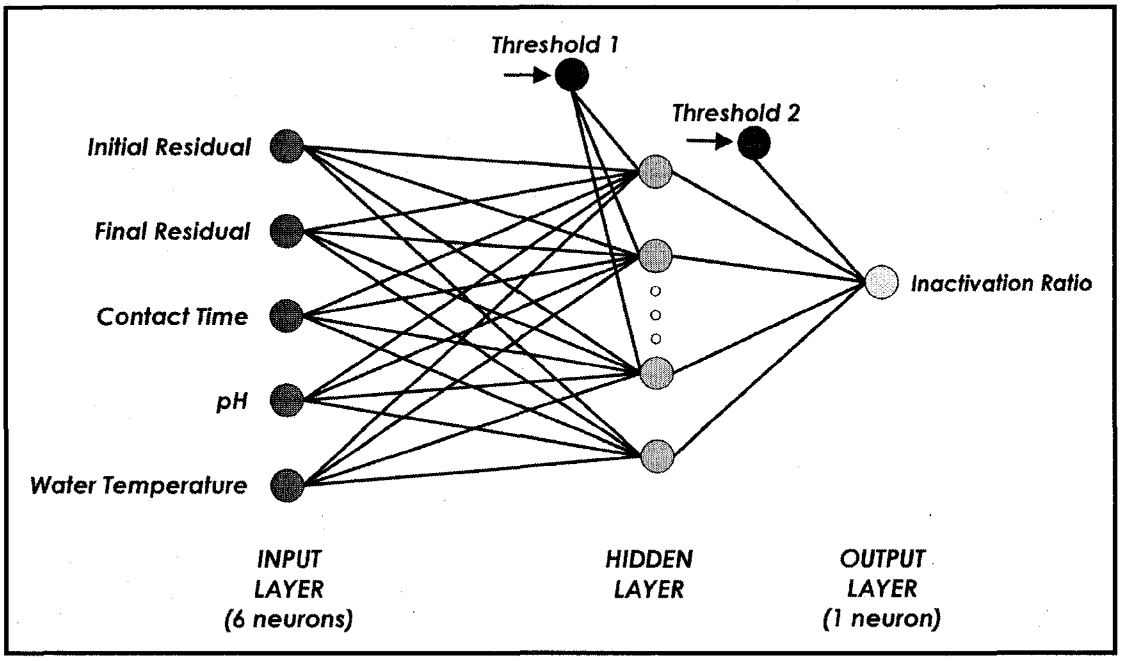

Figure 4.1: General neural network disinfection model architecture

4.3 Results and Discussion

4.3.1 Performance

The disinfection neural network models were trained with approximately 80% of available data and validated with the other 20%. The final trained network models, CDNN and ONN, and the Chick-Watson models, CW1 and CW2, were evaluated with the test sets for comparison. The performance results for all four models are available in Table 4.1. Only the test sets were used to evaluate the predictive capability of the neural network models. The training sets were not used for evaluating predictive capability since it is anticipated that the neural network models will have lower errors for the training data than the Chick-Watson models because the neural network models have several more elements that can be adjusted during model development. For the test sets the neural network and Chick-Watson models had marginally different R^2 values. However, for the test sets in terms of MAE the neural network models performed better than the competing Chick-Watson models. The MAE for CW1 was 10% more than the MAE achieved by CDNN, and the MAE for CW2 was over 25% higher than the error for ONN. The neural network models have performed well relative to the widely accepted Chick-Watson mathematical model, near-equivalent prediction capability has been observed. In order to definitively state that the neural network models have performed better than the Chick-Watson models it would be necessary to evaluate the predictive capability of the models using more test sets. There is currently no other data available for further evaluation and comparison.

	Test Set	
Model	**MAE**	**R^2**
CDNN (chlorine dioxide)	0.30	0.83
CW1	0.33	0.82
ONN (ozone)	0.31	0.87
CW2	0.39	0.88

Table 4.1: Disinfection model performance results

4.3.2 Single Input Saliency Analysis

The mean absolute errors were approximately double or greater for every omitted variable of both neural network models, indicating that all the input parameters are relevant to modeling these disinfection processes (Table 4.2). The input parameters for both models were ranked using MAE as the primary indicator and R^2 as a secondary measure.

Li *et al.* (2001a) found that the pH, for the 6 to 11 range, does not have a significant effect on chlorine dioxide inactivation, while temperature was vital for the inactivation of *C. parvum*. The omission of pH resulted in double the MAE, but was the second lowest increase and the coefficient of determination decreased marginally from 0.83 to 0.82, confirming that the CDNN model learned this trend. The mean absolute error almost tripled with the removal of temperature and the coefficient of determination decreased significantly to 0.25, demonstrating that CDNN considers temperature to be a significant factor for chlorine dioxide inactivation. Chlorine dioxide has a relatively stable residual, especially at low temperatures, thus the omission of the final residual was expected to generate the smallest increase in MAE and it did at 0.56. The most important input parameter for CDNN is the contact time because it produced an over threefold increase in MAE and the R^2 dropped to 0.11 when it was omitted. The input variables in terms of most too least relevant for CDNN are as follows: contact time, temperature, initial residual, pH, and final residual. Temperature was ranked to be important for the model than initial residual due to its dramatically lower coefficient of determination.

Li *et al.* (2001c) determined that at low temperatures the effectiveness of ozone to inactivate *C. parvum* decreases greatly, in other words, temperature is very important for ozone disinfection. For the ONN model, the removal of temperature yielded a R^2 of zero and approximately a fourfold increase in error, confirming the observation of Li *et al.* (2001c). Li *et al.* (2001c) found that pH from 6 to 8 is not a major influence on ozone inactivation of *C. parvum*. The ONN model also found pH to be less influential than other process variables but not completely irrelevant (Table 4.2). Interestingly, the least influential variables were found to be the initial and final residuals, while contact time was very significant. The disinfectant residual and contact time are generally considered to be equally essential for the inactivation of *C. parvum* (Li *et al.* 2001c). This suggests that the neural network models are selecting contact time as a feature to exploit in modeling the inactivation relationship, but could have equally selected the disinfectant residual as a more significant process variable. However, the omission of the final residual substantially increased the mean absolute error and decreased the coefficient of determination relative to the removal of the initial residual (Table 4.2), demonstrating the importance of including the final residual in the ozone disinfection model in order to account for the quick residual decay of ozone. The most influential parameters for ONN by far were temperature and contact time, followed by pH, final residual and initial residual.

Model	**Omitted Variable**	**MAE**	**R^2**	Rank
CDNN	None	0.30	0.83	
(chlorine dioxide)	Initial residual	0.86	0.71	3
	Final residual	0.56	0.72	5
	pH	0.66	0.82	4
	Temperature	0.84	0.25	2
	Contact time	1.03	0.11	1
ONN	None	0.31	0.87	
(ozone)	Initial residual	0.59	0.84	5
	Final residual	0.69	0.75	4
	pH	0.84	0.47	3
	Temperature	1.39	0.00	1
	Contact time	1.52	0.33	2

Table 4.2: Single input saliency analysis results using test sets

The back propagation based SLF algorithm found optimal network structures for the chlorine dioxide and ozone disinfection models: four and three hidden units respectively. The neural network models performed well relative to the temperature corrected Chick-Watson models through slightly lower errors produced and similar correlations between observed and predicted values for the test sets. Further model testing with new test sets is needed to determine whether or not the additional complexity of the neural network models over the temperature corrected Chick-Watson models are justified through better performance. For both neural network models the water temperature was an influential factor, while pH (for the ranges tested) was not as significant, which conforms to earlier studies (Li *et al.* 2001a; Li *et al.* 2001c). The inclusion of the final residual as an input parameter was very important for ozone, but not as critical for chlorine dioxide. It seems the less stable a disinfectant residual is, the more important the final residual is to a neural network disinfection model's prediction performance. The incorporation of the final residual into a neural network disinfection model is an effective first step in addressing the concern identified by Haas (2004) of how to consider the disinfectant residual decay within a neural network model.

5.0 WATER CONSUMPTION MODEL

The World Health Organization has identified the most prevalent health risk for drinking water to be the infectious diseases caused by waterborne pathogens, such as *Cryptosporidium parvum* (WHO, 2004). In Canada, tap water accounts for a large portion of daily fluid intake and is a main concern for transmission of waterborne pathogens (CEHD, 2001; WHO, 2004). The neural network consumption models developed in this study address the consumption of tap water as the exposure route for a waterborne pathogen. The data used to develop the neural network consumption models are from a 1970's Canadian nationwide survey.

Many exposure assessment studies for waterborne pathogens only consider the total water consumed, and do not differentiate between the amounts of hot and cold water consumed (Ershow and Cantor, 1989). While total tap water consumption might be appropriate for chemical risk assessments, this approach is often not suitable for microbial risk assessment. When considering source data for the consumption models in this research, it was necessary to consider the amount of hot tap water consumption, since heating water deactivates many pathogens, including *C. parvum.* Jenkins *et al.* (1997) report complete loss of infectivity for *C. parvum* oocysts that are exposed to temperatures of 60°C or greater for 5 minutes. This is a very significant consideration because the

Canada Safety Council (2005) reports that most Canadian homes set their hot water heaters at 60°C or higher. Generally, the amount of cold tap water a person consumes during a specific time period is used to determine their exposure to a pathogen, given the concentration of the pathogen in the tap water. It is very important that risk assessors have accurate cold tap water consumption rates in order to provide realistic risk estimates for infection and illness when exposure to contaminated water occurs. In microbial risk assessment total tap water consumption should be considered if the risk assessor would like to be conservative, or if it is believed the water may not have been heated sufficiently.

It is often the practice in risk assessment to use a single point estimate for the consumption of water to represent daily exposure for large segments of the population (WHO, 2004). The World Health Organization assumed a per capita daily consumption of 1 litre of cold water to develop its guidelines for microbial hazards (WHO, 2004). This assumed daily consumption is very high for children, and very low for older adults, which may overestimate and underestimate the risk for these groups through quantitative risk assessment. The United States Environmental Protection Agency (EPA) uses an average consumption rate of 1.41 litres of total tap water per day for adults, and makes no adjustment for hot tap water (OEHHA, 2000). The neural network consumption models developed in this research provide average daily consumption rates for total tap water and cold tap water. There are many factors that the point estimate from the WHO does not consider, such as age and gender. Various studies have shown significant variations in tap water consumption by age group, gender, season, and geographic location (CEHD, 1981; Ershow and Cantor, 1989; DWI, 1996; Roseberry and Burmaster, 1992). These parameters have been included in the neural network tap water consumption models developed in this research.

5.1 Datasets

The tap water consumption models developed for this research are targeted at Canadians and are based on data from a Canadian Ministry of National Health and Welfare study on tap water consumption conducted in the summer of 1977 and winter of 1978. This government study involved 970 people from 295 households representing people from across five regions of Canada (British Columbia, the Prairie Provinces, Ontario, Quebec and the Maritimes) (CEHD, 1981). The study only considered tap water consumption; however consumption was broken down into cold and hot beverages. This is very important for the consumption of many waterborne pathogens because many pathogens are not able to survive even short periods of extreme heating. The raw information for the study was collected using questionnaires and interviews. The results of the survey were analyzed in terms of age, gender, season, and geographical location. Similar studies performed around the same period in Holland, the United Kingdom, and the United States produced similar intake patterns

(OEHHA, 2000; CEHD, 1981). Despite the age of this study, the data is still considered to be representative of the tap water consumption for Canada and has recently been used by the American Environmental Protection Agency (EPA) to estimate American tap water consumption (OEHHA, 2000).

The Drinking Water Inspectorate (DWI) of the United Kingdom compared a more recent 1995 tap water consumption study in England and Wales with a 1978 national survey in the UK performed by the Water Research Center and made several relevant conclusions. First, the total liquid consumption did not change significantly from 1978 to 1995; and second, the consumption of bottled water has increased significantly, while the consumption of tap water was not affected by this increase (DWI, 1996). Other beverages such as soft drinks have experienced declines in consumption rates. Canada and the UK have experienced similar consumption patterns in the past and it is assumed that this trend has continued. It is assumed that the tap water consumption patterns observed in 1977 and 1978 during the Canadian survey are representative of the intake patterns for Canadians in 2007.

Two separate and distinct datasets available from the Canadian survey were used to develop and test two neural network consumption models that produced point estimates for water consumption. The first dataset had information on the season (winter versus summer), as well as the age and gender of the consumer. This dataset was used to develop a model that was termed the Seasonal Neural Network (SNN). The second dataset contained information on five geographic regions across Canada (British Columbia, the Prairie Provinces, Ontario, Quebec and the Maritimes), and the age and gender of the consumer. This dataset was used to develop the Regional Neural Network (RNN) consumption model. Both datasets included the total tap water consumption and cold tap water consumption. Both datasets were analyzed for outliers and missing data. All inputs and outputs for data patterns were found to be within ±3 standard deviations and there was no missing data.

The allocation of data between training sets and testing sets was approximately 75% and 25% respectively for the developed models. The test points were randomly selected from the complete datasets. The arithmetic mean of the input and output variables for the training and testing sets had to be within 50% of each other. This was done to ensure that the data in the training and test sets were reasonably representative of each other and still randomly selected. A total of 54 data points were available for developing and testing the SNN model and 108 points for the RNN. For each age group the SNN dataset contains a discrete point for each of the following combinations:

- Male/summer, female/summer, and male/female/summer;

- Male/winter, female/winter, and male/female/winter, and;
- Male/both seasons, female/both seasons, and male/female/both seasons.

For each age group the RNN dataset contains a discrete point for similar combinations of male, female, combined genders, individual regions, and the entire country. However, the RNN dataset does not provide discrete points for combinations of age groups or regions (beyond the entire country). Thus the RNN is expected to generalize the water consumption relationship so that a point estimate for adult males (over the age of 19) in Quebec and Ontario could be produced for example.

5.2 Neural Network Consumption Models

The neural network consumption models have been trained using SLF, tested using segregated test sets and examined using saliency analysis. The first developed neural network, SNN, includes the season, summer or winter, as an input. The second model, RNN, uses regions within Canada as an input instead of seasonal information. Both neural networks have age and gender as input variables, and total tap water consumption and cold tap water consumption in litres per day as output neurons.

The consumption of tap water is often assumed to be equivalent for immunocompetent and immunodeficient populations (Pouillot *et al.*, 2004).

However, the immunodeficient population is often more susceptible to waterborne illness than the immunocompetent population, and decreases their tap water intake opting for bottled water instead (Aragon *et al.* 2003). Assuming that the tap water intake rates are similar will likely overestimate the risk to the immunodeficient populations, which provides a precautionary element. This study assumes that the consumption of tap water is equivalent for immunocompetent and immunodeficient populations.

5.2.1 Seasonal neural network consumption model

The seasonal neural network model, SNN, included gender, age and season as input variables. These variables were selected because previous studies indicated that these factors are important contributors to consumption rates, with age being the most significant (DWI, 1996; OEHHA, 2000; Ershow and Cantor, 1989). There are a total of 10 binary inputs to the network, where one indicates the input is active and zero indicates inactive. Gender is represented as two binary inputs for male and female; when both the male and female inputs are active this indicates a combination of both sexes. Age is represented by six binary inputs for six different age categories in terms of years old: less than 3, 3 to 5, 6 to 17, 18 to 34, 35 to 54, and 55 and over. Activating different age inputs at the same time will produce point estimate consumption rates for different combinations of age groups. The season is represented by two binary inputs for summer and winter; activating both inputs represents combined seasons or the entire year. It is assumed that the summer season is representative of the summer and spring, while the winter is representative of fall and winter.

The network has been trained with data representing the daily consumption rates for males, females, and combined genders in every age category for the summer, winter, and combined seasons. The network model has not been trained with data representing combinations of age groups, such as 3 to 5 and 6 to 17 simultaneously. However, the SNN is theoretically capable of providing average consumption rate estimates for combinations of age groups due to the generalization capability of neural networks.

The SNN model started from a large fully connected structure with 10 binary inputs, 7 hidden neurons, and 2 output neurons and was trained using back propagation based SLF. In response to the test set, the trained SNN model achieved a MAE of 0.0345 and R^2 of 0.986 for total tap water consumption, and a MAE of 0.0355 and R^2 of 0.940 for cold tap water consumption.

The connection weights of the SNN structure were examined and weights less than $1x10^{-4}$ were considered redundant. If all the weights connected to a particular neuron (input, hidden, or output) were less than $1x10^{-4}$, the neuron was also considered redundant. This policy resulted in four hidden neurons and the two seasonal inputs, summer and winter, being considered redundant, and indicated that for Canadian tap water consumers the season is irrelevant. Saliency analysis was performed to determine the relative importance of each input variable and if the seasonal input actually hindered the model's performance.

The saliency analysis was performed using the same 14 data point test set; the particular input variable being tested was omitted by zeroing all the related input neurons. For example, when age was tested all six input neurons representing the six age categories were zeroed in the test set. The results of the saliency analysis for SNN are presented in Table 5.1. The MAE for both total and cold tap water consumption increased significantly with the omission of age, doubled with the removal gender, and decreased slightly with the omission of the seasonal inputs. These results indicate that age is the dominant factor as expected in determining tap water consumption, while gender is a relatively minor factor. The inclusion of seasonal inputs slightly degraded the model's performance, and should be excluded from the model. It is generally thought that seasonal temperature differences would affect tap water consumption. However, it is possible that the season affects total fluid consumption instead of tap water consumption, which only represents a fraction of overall fluid intake.

Variable Omitted	MAE Total Tap Water	R^2 Total Tap Water	MAE Cold Tap Water	R^2 Cold Tap Water	Rank
None	0.0345	0.986	0.0355	0.940	
Age	0.273	0.247	0.141	0.129	1
Gender	0.0602	0.978	0.0674	0.800	2
Season	0.0321	0.987	0.0354	0.945	3

Table 5.1: Seasonal neural network model (SNN) results and saliency analysis

5.2.2 Regional neural network consumption model

The second developed neural network, the regional neural network (RNN) model, has three input variables: gender, age, .and regions within Canada. There are a total of 13 binary inputs and 2 outputs for the model. The gender and age inputs are similar to the SNN model. For the regions, there are five binary inputs representing the Maritimes, Quebec, Ontario, the Prairies, and British Columbia. When all five regions are active, this is representative of the entire country. The architecture for the RNN is presented in Figure 5.1.

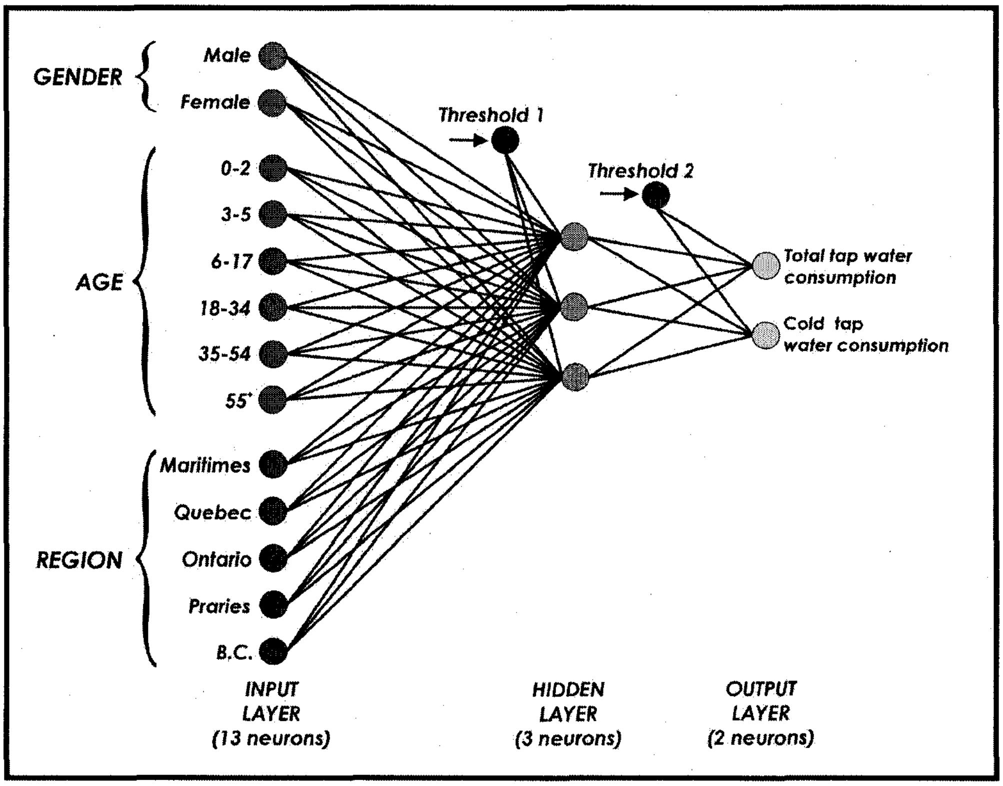

Figure 5.1: Regional neural network (RNN) consumption model

The network has been trained with data representing the daily consumption rates for males, females, and combined genders in every age category for each region and the entire country. The network has not been trained with data for combinations of regions other than the overall country, but is hypothetically capable of generating such estimates due to the architecture design and generalization capability of neural networks. Also, similar to the SNN, the RNN model is theoretically capable of providing estimates for different age group and gender combinations.

The RNN model was trained using SLF with an initial structure of 10 hidden neurons, which was reduced to 3 hidden neurons after training and application of the policy that connection weights less than $1x10^{-4}$ are eliminated. None of the input neurons were eliminated through training. In order to verify that 3 hidden neurons was the optimal structure in terms of MAE and R^2, the number of hidden neurons was varied from 2 to 6 through a trial and error approach. The penalty term for SLF was set to zero during this trial and error validation, reducing the learning algorithm to regular back propagation learning. It was confirmed that three hidden neurons was the optimal structure. The final RNN model was trained with SLF, with a three hidden neuron structure and produced the smallest MAE of 0.0184 and 0.0173 for total and cold tap water consumption respectively. The RNN model greatly outperformed the SNN model in terms of MAE and showed slight improvements in R^2 values. The results of the RNN model are presented in Table 2.2. The WHO recommends using 2 litres of total water per day and 1 litre of cold water per day as point estimates for consumption rates of tap water (WHO, 2004). These WHO point estimates produced an MAE of 0.787 and 0.313 for total and cold tap water consumption rates respectively against the 27-point test set used for testing the RNN. The MAE values for the WHO point estimates are approximately 42 and 18 times higher than the MAE values of the RNN model for total and cold tap water consumption respectively. This result demonstrates the need for including factors such as age in consumption rate estimates, instead of using a single point estimate to represent the entire population.

Saliency analysis of the RNN model was carried out using the same test set used to validate the model after training. The saliency analysis results are presented in Table 5.2. As expected, the MAE values increased significantly and R^2 values decreased when age was omitted. The omission of gender from the model approximately tripled the MAE values and slightly decreased R^2 values; the omission of regions produced similar results. Age is the dominant factor in determining tap water consumption rates, while gender and regions are approximately equal in importance for the RNN model.

Variable Omitted	MAE Total Tap Water	R^2 Total Tap Water	MAE Cold Tap Water	R^2 Cold Tap Water	Rank
None	0.0184	0.997	0.0173	0.987	
Age	0.360	0.127	0.1267	0.255	1
Gender	0.0503	0.972	0.0663	0.855	2
Region	0.0645	0.957	0.0595	0.861	2

Table 5.2: Regional neural network model (RNN) results and saliency analysis

In order to analyze the internal organization of the neural network, the training set and test set for the RNN model were combined, for a total of 108 data points. Cluster analysis of the hidden unit responses to each input pattern was performed. An interesting regularity was discovered by the neural network model and used to structure the knowledge it acquired.

Hidden neuron one detected the composition of total tap water consumption between cold and hot tap water. In the first three age groups (less than 3, 3 to 5, and 6 to 17 years of age) for both genders the amount of cold tap water represents a large portion, between 70% and 90%, of total tap water consumption. In comparison, for the three older age groups (18 to 34, 35 to 54, and 55 and over) cold tap water accounts for only 30% to 60% of total tap water consumption. This difference can be accounted for through higher rates of coffee and tea drinking among adults (18 and over) relative to children, and higher rates of frozen concentrate juice consumption among children and adolescents. Hidden neuron one would produce an intermediate activation (approximately 0.5) for adults (18 and over) and deactivate for children and adolescents.

5.3 Discussion

It was determined that age, region and gender are relevant variables in the determination of water consumption for Canadians and that seasonal differences are marginal. Thus, the RNN consumption model will be used to determine the cold tap water consumption of Canadians for the exposure assessment of waterborne *C. parvum* in this thesis.

There are other factors that can affect tap water consumption rates, such as physical activity level and community size. As individuals perform more strenuous tasks their water consumption rate will generally increase to account for fluid losses to perspiration. The difference in total daily tap water consumption rates between rural areas and communities over 500,000 people is considered appreciable (CEHD, 1981). However, there was insufficient data available to include these variables in the developed neural network consumption models. Inclusion of these variables in future consumption models would provide even greater segmentation of the population for generating very specific point estimates of water consumption.

6.0 DOSE RESPONSE MODEL

6.1 Literature Review

Quantitative microbial risk assessment offers modeling methods that combine exposure assessments and dose response relationships to produce estimates of population health effects for a given microbial pathogen (WHO, 2004). A dose response model is central to the performance of a QMRA and provides the probability of infection after exposure to one or more pathogens. The median infectious dose represents the dose at which 50% of the population will become infected (Thomas and Hrudey, 1997).

It can be very difficult to generate an accurate dose response relationship due to a lack of data at the low, intermediate and high dose levels, and due to the high variability of host susceptibility across the population (Beaglehole *et al.*, 1993). In order to overcome these problems, DuPont *et al.* (1995) performed *Cryptosporidium* challenge studies of healthy adult human volunteers using a wide range of doses in order to develop a median infective dose and dose response relationship for the *C. parvum* pathogen in its Iowa isolate. The results of the study show that dosages of 300 oocysts or more produced an 88% infection rate, and the median infective dose was found to be 132 oocysts for the Iowa strain (DuPont *et al.*, 1995). Therefore, in healthy adult humans a relatively low dose of *C. parvum* oocysts is enough to produce infection. Since this study was performed using human subjects instead of animals, this evidence is considered to be very strong for representing the infectious response of healthy adults with no immunity for the Iowa isolate of *C. parvum.*

Similar *Cryptosporidium* challenge studies on healthy adult volunteers were performed for the TAMU and UCP strains of *C. parvum.* The median infectious doses for the TAMU and UCP strains were found to be 12 and 2066 oocysts respectively (Messner *et al.*, 2001). The TAMU and UCP studies were performed with smaller groups of 14 and 17 people respectively with doses ranging from 10 to 500 oocysts for the TAMU strain and 500 to 10 000 for the UCP strain (Messner *et al.*, 2001). Since these three *Cryptosporidium* challenge studies were performed using healthy adult volunteers it is very likely that the median infectious dose is lower for high risk groups such as Acquired Immune Deficiency Syndrome (AIDS) patients.

The three studies for the Iowa, UCP and TAMU strains were analyzed by Messner *et al.* (2001) with the assumption of an exponential dose response relationship. The exponential model is a single parameter model that does not have a minimum infectious dose. It has been shown that a single oocyst can cause infection (Messner *et al.*, 2002).

The susceptibility of a host becoming infected and ill from *C. parvum* is directly related to the amount of exposure to the pathogen. It is also dependent on a number of other variables including the immune competence of the host, the strain of *C. parvum* exposed to, and recent *C. parvum* infections within the last year. These factors were not considered by the exponential dose response models developed by Messner *et al.* (2001) for the Iowa, UCP, and TAMU strains.

When a healthy immunocompetent adult becomes infected with *C. parvum* and becomes symptomatic, their symptoms are typically present for less than four weeks, however the excretion of oocysts can persist up to eight weeks (Martins and Guerrant, 1995). A symptomatic infection for an immunodeficient person (characterized by a low T-cells count) can have far more devastating effects (Hrudey and Hrudey, 2004). AIDS patients and others who are immunodeficient, such as chemotherapy patients, can develop persistent diarrhea that can result in death.

The three *C. parvum* challenge studies reflect the varying infectivity of *C. parvum* strains. The TAMU strain is 12 times more infective than the IOWA strain and 172 times more infective than the UCP strain given the median infective doses. This large variation of infectiveness between the strains introduces a lot of uncertainty for risk assessment should assessors assume that all oocysts in an outbreak are of a particular strain or a combination of the strains. The single parameter exponential models developed by Messner *et al.* (2001) do not provide the ability to consider combinations of strains.

Hosts infected by C. *parvum* develop *Cryptosporidium* antibodies and gain a degree of protection against being re-infected. An analysis study by Teunis *et al.* (2002) of data from the DuPont *et al.* (1995) *C. parvum* study shows that this immunity is short-lived, lasting less than 12 months. This suggests that as an outbreak progresses the number of people who develop this short-term protection will increase and the outbreak will eventually peak as the number of susceptible hosts decreases.

It is difficult to incorporate these variables as inputs to a conventional statistical model. However it is comparatively straightforward for a neural network model. The neural network dose response model developed for this study considers the AIDS status, multiple strains of *C. parvum*, and immunity for the IOWA strain as input variables.

In the 1993 Milwaukee *C. parvum* outbreak, which affected approximately 400,000 people, it was found that the attack rates were lower in the younger age groups (MacKenzie *et al.*, 1994), which is counter intuitive since it is usually assumed that young children will be more susceptible to infection than healthy adults. Perz *et al.* (1998) developed a model that assumed that the infectivity was the same for children and adults. This research followed the precautionary principal and assumed that the infectivity of *C. parvum* is the same for all age groups.

6.2 Dose Response Datasets

The datasets used to develop the neural network dose response model are from the available literature on several separate *C. parvum* challenge studies using healthy adult volunteers. A study by DuPont *et al.* (1995) used the Iowa isolate and involved 29 adult volunteers between 25 and 35 years of age who were given single doses of oocysts at eight different dosing levels ranging from thirty to one million *C. parvum* oocysts. These volunteers were tested for the lack of *Cryptosporidium* antibodies to ensure that protective immunity would not be an issue. A separate study by Chappell *et al.* (1999) selected 17 healthy adult volunteers based on the presence of *Cryptosporidium* antibodies, which indicates a recent *C. parvum* infection within the last year, which builds some immunity to the pathogen during subsequent exposures. Four different dosing regiments were used in that study: 500, 5000, 10 000, and 50 000 Iowa strain oocysts. Dr. Harley Moon of the University of Iowa collected the original isolate used in these studies from a calf.

The other *C. parvum* challenge studies performed by Chappell *et al.* (1999) only used healthy adult volunteers who had not been exposed to *C. parvum* within the last year and showed no immunity. The UCP isolate, also originally derived from a calf, was given to 17 different subjects in doses of 500, 1000, 5000 and 10 000 oocysts. The TAMU isolate, originally derived from an infected veterinarian student, was given in doses of 10, 30, 100 and 500 oocysts to 14 volunteers.

The increased sensitivity of the AIDS population to infection was considered in the developed *C. parvum* dose response neural network model. There is currently no experimental data available on the infectivity of *C. parvum* among the AIDS population, such an experiment would be highly unethical, thus available outbreak data must be replied upon. Perz *et al.* (1998) estimated that the infectivity for *C. parvum* for the AIDS population was three times higher than the non-AIDS population based on available *C. parvum* outbreak data. Makri *et al.* (2004) applied a factor of three to represent the increased infectivity among the AIDS population for their risk assessment model based on New York City data. To account for AIDS status in the neural network dose response model, a factor of three was applied to the Iowa, UCP and TAMU dose response relationships produced by the neural network model after training with the available experimental data for the three strains. The dose response relationship for the Iowa strain with built up immunity among subjects was not considered for the AIDS population since once infected, AIDS patients are assumed to develop illness and never recover. Pouillot *et al.* (2004) also assumed certain illness among infected AIDS patients in their risk assessment of *C. parvum* in the French population.

The dataset used for training the neural network model is presented in Table 6.1. This table does not include the data used for those populations with AIDS; this is simply the data for the three isolates whose receptors have no previous immunity, multiplied by a factor of three to be protective of sensitive populations with AIDS.

Strain of *C. parvum*	Dose (Oocysts)	Number of Exposed	Number of Infected
Iowa	30	5	2
	100	8	4
	300	3	2
	500	6	5
	1 000	2	2
	10 000	3	3
	100 000	1	1
	1 000 000	1	1
TAMU	10	3	2
	30	3	2
	100	3	3
	500	5	5
UCP	500	5	3
	1 000	3	2
	10 000	4	4

Table 6.1: *C. parvum* dose response dataset for model development

For the UCP strain there was a fourth data point at a dosing level of 5000 oocysts that was considered extraneous and eliminated because it did not follow the trend of the other data patterns for the UCP. No data patterns were removed for the other strains.

Another *C. parvum* challenge study was performed by Okhuysen *et al.* (2002) using a fourth isolate, Moredun. In Okhuysen *et al.*'s (2002) study, 16 adult volunteers received a dosing regiment ranging from 100 to 3000 oocysts. This dataset was not used to develop the neural network dose response model because the results of the study failed to demonstrate a significant relationship between the ingested dose and the onset of infection for the applied dose range.

6.3 Exponential Dose Response Models

The exponential dose response model is a very simple one-parameter model that assumes there is no minimum infectious dose. It has been demonstrated through animal and tissue models that even a single oocyst can cause infection (Messner *et al.*, 2001). There have been no human studies conducted to date that can demonstrate a single oocyst causing infection in a human subject. Messner *et al.* (2001) used the exponential dose response model for their models of the Iowa, TAMU and UCP strains:

$$P(D,k)=1-e^{-D/k}$$

where the probability of infection, *P*, is a function of the dose, *D*, in terms of the number of oocysts, and a constant dose response parameter, *k*. Messner *et al.* (2001) used the maximum likelihood parameter estimate to determine that the value of k was 17.5 for the TAMU isolate, 190 for the Iowa isolate, and 2980 for the UCP isolate. In order to compare the exponential models against the developed neural network dose response model it was necessary to determine a new value for the parameter *k* for the UCP strain because the data available for the 5000 oocyst dose level was determined to be extraneous. Using the same method as Messner *et al.* (2001), a new maximum likelihood estimate was calculated for the UCP strain with dose response information at 500, 1000, and 10 000 oocysts, and the new constant dose response parameter *k* was 546.

6.4 Neural Network Dose Response Model

A three-layer MLP neural network with gradient-descent based backpropagation training was used to develop one neural network dose response model for *C. parvum*. The input layer of the model considered five inputs: the log doses for the Iowa, UCP, and TAMU strains of *C. parvum*, AIDS status, and previous infection from the Iowa strain within the last year. The output layer of the model has one output unit representing the probability of infection assuming exposure. The occurrence of diarrhea and other complications after infection were not considered in the model due to lack of available information for model development. The optimal number of hidden neurons for the final neural network dose response model was determined to be two. The architecture for the neural network dose response model has been presented in Figure 6.1. The neural network disinfection model was developed in C++ and trained on an IBM compatible computer with a 2.93GHz Intel Celeron processor and 1GB of memory.

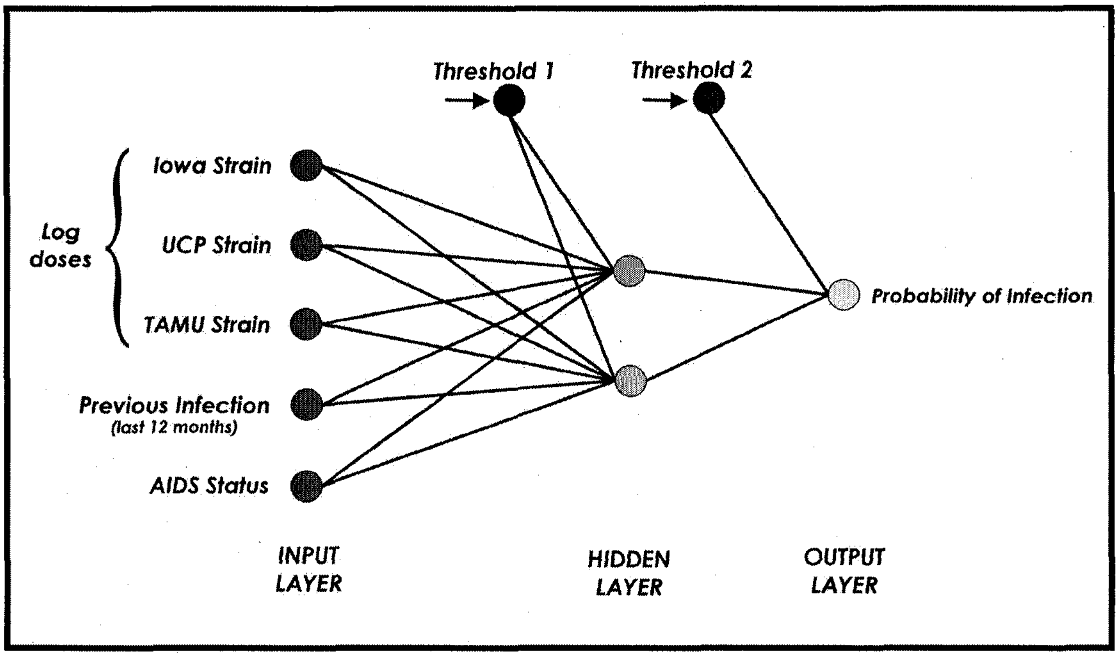

Figure 6.1: Dose response neural network model architecture

There was no low dose data (1 oocyst) available for training the neural network, the same problem encountered by statistical models. The exponential model assumes that there is no minimum infectious dose. For the neural network model, it was assumed that a single oocyst is capable of causing infection within a human subject, and thus the probability of infection for log dose of zero (one oocyst) must be greater than one for the dose response model. It was assumed that the probability of infection was zero when the log dose was equal to –7. This assumption was included in the training set for the neural network for each of the three strains of *C. parvum* considered. Ce Yu (2004) made a similar assumption for a knowledge-based neural network dose response model of E. coli O157:H7. If a risk assessor wants to be more or less cautious they should adjust accordingly the selection of what the log dose equals for the probability of infection to be zero. The incorporation of human knowledge at the low dose was necessary to train the neural network model in the absence of low dose response data, however when low dose data becomes available it should be incorporated into the neural network dose response model.

6.5 Results and Discussion

The neural network dose response model was capable of fitting all the data available for the three strains of *C. parvum*. For each strain of *C. parvum* the neural network dose response model achieved higher correlations and lower errors than the corresponding exponential models, the results are presented in Table 6.2. However, performance gains were not significant in many cases, which lead to the argument that the complexity of the neural network model should be abandoned for the simplicity of the exponential model. In principal, the developed neural network model has the capability of predicting the frequency of infection for a water sample containing each of the three strains of *C. parvum* under study.

Model	Strain of *C. parvum*	R^2	MAE
Exponential	Iowa	0.955	0.0634
	TAMU	0.806	0.0970
	UCP	0.700	0.102
Neural Network	Iowa	0.974	0.0401
	TAMU	0.866	0.0627
	UCP	0.879	0.0784
	Iowa (previous exposure)	0.874	0.115

Table 6.2: Dose response model results

The dose response relationships for each of the three strains are presented in Figures 6.2 to 6.4. Each graph presents the exponential and neural network dose response relationships along with the observed responses used for training. The Iowa graph, Figure 6.2, also depicts the dose response relationship for individuals that were previously infected by the strain within the last year. The previous infection clearly provides some level of immunity since the frequency of infection dropped significantly at each dose level. The TAMU graph, Figure 6.4, also depicts the dose response relationship for the AIDS population. The low dose data point based on available knowledge has not been included in any of the graphs. As expected each of the neural networks' dose response relationships are higher than the corresponding exponential model in the low dose range. This higher response provides better protection for human health in the absence of real human response data in the low dose range.

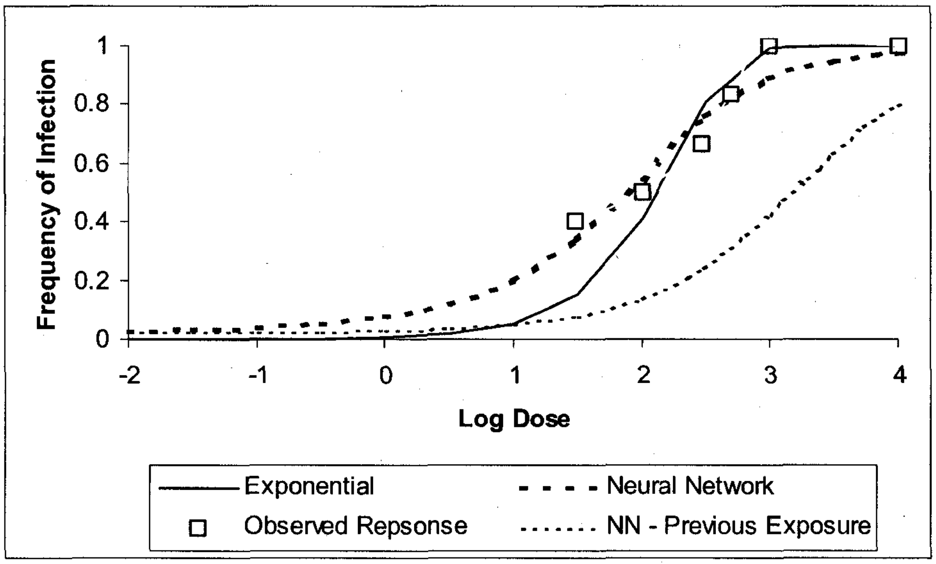

Figure 6.2: Iowa strain dose response relationship

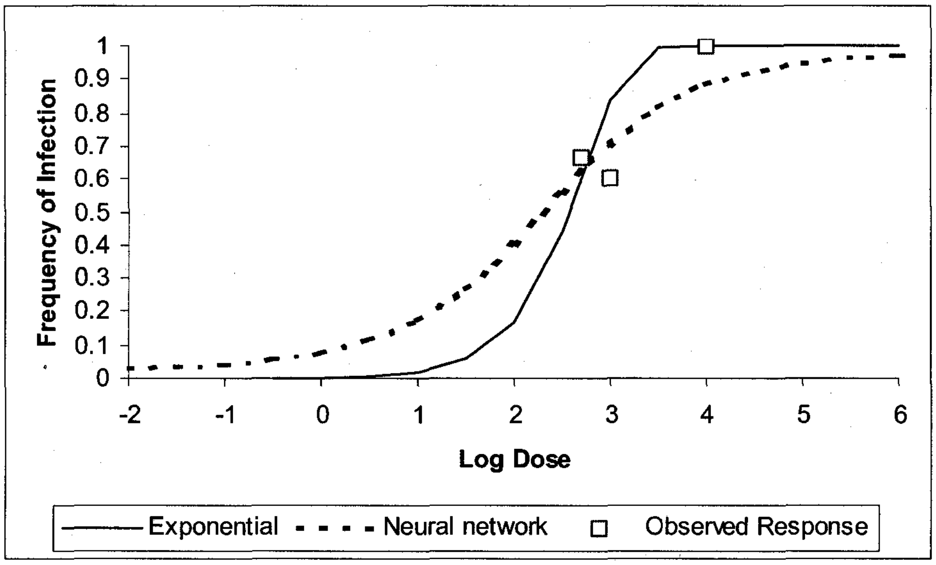

Figure 6.3: UCP strain dose response relationship

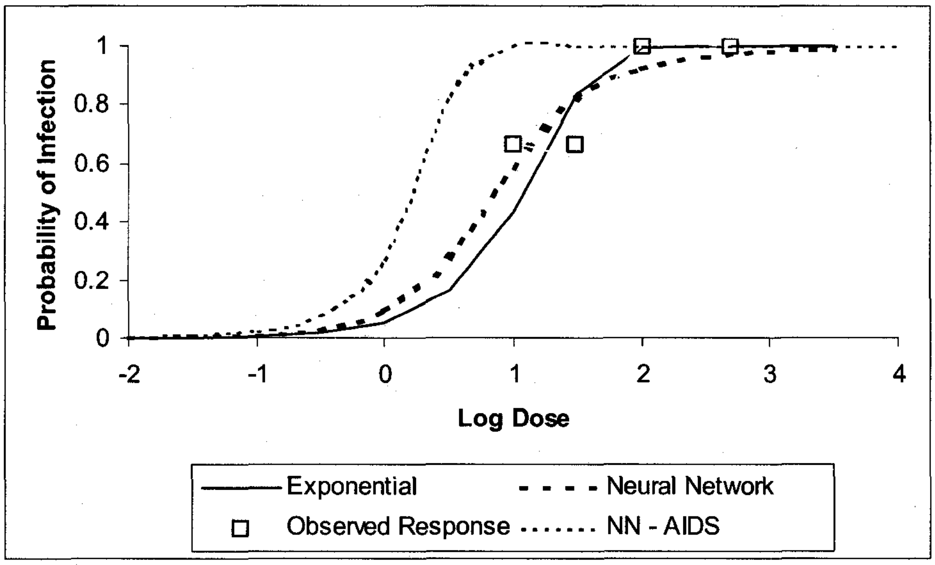

Figure 6.4: TAMU strain dose response relationship

This neural network dose response model provides risk assessors with an excellent tool in the quantitative microbial risk assessment of *C. parvum* in drinking water. The ability to simulate different combinations of strains that have different levels of infectivity could be useful when the exact strain of *C. parvum* exposure is unknown.

7.0 RISK CHARACTERIZATION

Risk characterization integrates the outputs of the dose response and exposure assessments into a quantitative risk estimate (Haas *et al.*, 1999). It is also necessary in the risk characterization to decide which output measure will be produced by the risk assessment, such as the expected number of illnesses for a population, or the probability of infection. The final step of this risk assessment process is to combine the water consumption per day, oocyst concentration in the drinking water and the dose response relationship to determine the probability of infection from *C. parvum*.

There is a significant decision made at the outset of a risk assessment that greatly affects the output of the risk characterization; the decision to use point estimates of risk or interval estimates. The expected number of illnesses for example is a point estimate of risk since it is a single numerical value representing risk. An interval estimate of risk is given as a probability distribution or confidence region, and thus considers the uncertainty and variability of the inputs and assumptions (Haas *et al.*, 1999). A point estimate of risk for exposure to *C. parvum* via drinking water is calculated by determining a point estimate of the oocysts consumed through cold drinking water from a single exposure (one day) and inputting that point estimate of exposure into the dose response model.

Point estimates of risk are straightforward to compute and easy to understand relative to interval estimates of risk. However, the simplicity of point estimates of risk sacrifice the ability to analyze uncertainty and variability within the developed models. In this thesis point estimates of risk were produced from the risk assessment process and interval estimates of risk were not considered.

7.1 Risk Characterization Assumptions

There are endless possible combinations of parameters to consider with the exposure and dose response assessments for the risk characterization. The purpose of the risk characterization in this application is to demonstrate the use of the developed neural network models together. The simulations performed consider variations in the oocyst concentration, disinfection process, drinking water consumption, and dose response parameters including strain, previous exposure, and AIDS status. A total of 108 different simulations were conducted to determine point estimates of daily probabilities for infection.

7.1.1 Oocyst Concentrations

Oocyst concentrations of 12, 119 and 250 oocysts per liter were assumed for the source waters before treatment. This provides a range of raw water source qualities for the simulations, from good to terrible, which can be expected depending on the land uses for the surrounding watershed.

7.1.2 Disinfection Process

For the disinfection process relatively low, medium and high inactivation rates were desired for the risk characterization in order to simulate a failure of the disinfection process and when the process is operating properly. The ozone disinfection neural network model was selected for this purpose, as only one disinfection process was required for simulation purposes. It was necessary to use three test points used to evaluate the ozone disinfection model for the risk characterization simulations because the model requires the final residual as an input and it was not possible to generate additional data points (in an operational plant the final residual could easily be measured). The model inputs and the neural network inactivation rate outputs for the three test data points used for risk characterization simulation are presented in Table 7.1.

Initial Residual (mg/L)	Final Residual (mg/L)	Contact Time (minutes)	pH	Temperature (°C)	Neural Network Inactivation Rate (log units)
1	0.7	15	8	1	0.27
1.8	1.5	15	6	1	1.45
0.5	0.3	20	6	22	3.44

Table 7.1: Water consumption model inputs and inactivation rates

7.1.3 Drinking Water Consumption

Drinking water consumption was considered for two different groups, Canadian adults (18 and over) and children (under 18). Although the developed neural network water consumption model is capable of considering different genders, regions within Canada, and a variety of age groups; age was determined through saliency analysis to be the most important variable considered by the model. Thus it was decided that age would be the driving consideration for the consumption of water for the simulations performed. The two age groups were selected based on an observation from the consumption model that adults drink considerably less cold tap water as a portion of their total tap water consumption than children (under 18).

It is assumed for this study that all Canadians keep their hot water heaters at 60°C or above, and that the water will be heated in the tank for at least 5 minutes. Otherwise it is assumed that Canadians will sufficiently heat their hot water drinks to kill all present oocysts. Thus hot tap water consumption does not need to be considered for this exposure assessment and only cold tap water consumption needs to be considered. The cold tap water consumption for Canadian adults and children was determined by the regional tap water consumption neural network model to be 0.97 and 0.64 liters per day respectively.

7.1.4 Dose Response Parameters

The following separate *C. parvum* dose response relationships from the dose response neural network model were considered during the simulations: UCP and TAMU strains assuming independent exposures, Iowa strain assuming no previous exposure, Iowa strain assuming a previous exposure within the last 12 months, and the Iowa strain for an AIDS positive person assuming independent exposures. While the dose response neural network has been trained with the consideration of AIDS status for the TAMU and UCP strains, it was deemed that demonstration of this condition for one strain was sufficient.

7.1.5 Exposure Independence

A population can be exposed to a pathogen multiple times or continuously. For most pathogens there is no experimental data surrounding multiple or continuous exposures, and it is assumed that the risk of each exposure is statistically independent of the risk from other exposures. However, in the case of the Iowa strain for *C. parvum* there is multiple exposure data available, which was included in the developed dose response neural network model. For this risk characterization it is assumed for UCP and TAMU strains of *C. parvum* that each exposure is statistically independent. Temporary or permanent immunity reduces the independence of a series of exposures, but the assumption of independence of exposure is considered to be reasonable (Haas *et al.*, 1993). For the Iowa strain it is assumed that in the case of multiple exposures, the dose response relationship for the Iowa strain given previous infection within 12 months will be applied.

7.2 Simulation Results

The simulation results for source water oocyst concentrations of 12, 119 and 250 oocysts per litre are presented in Tables A.1 – A.3 of Appendix A. First consider the best-case scenario with the highest rate of inactivation (99.96%), lower cold tap water consumption (Canadian children at 0.64 L per day), and the lowest considered oocyst concentration at the water source (12 oocysts per L). This scenario produced a daily probability of infection ranging from 0.016 to 0.034 for the Iowa, UCP, and TAMU strains with no previous infection. It is necessary to remember that the developed neural network dose response model is very conservative in the low dose range (1 oocyst or less). Now consider the worst case: low rate of inactivation (46.29%), adult cold tap water consumption (0.97L per day), and the highest oocyst concentration for source water considered (250 oocysts per L). This situation generated a daily probability of infection ranging from 0.343 to 0.989 for the Iowa, UCP and TAMU strains with no previous exposure. Clearly, the difference in the probability of infection is substantial between the two scenarios. The lowest rate of inactivation (46.29%) used in the simulations was meant to represent a poorly operated or failing ozone disinfection process. The ramifications of a failing disinfection process are quite evident from the relatively high probabilities of infection for all scenarios involving a low rate of inactivation.

Previous infection to the Iowa strain of *C. parvum* in the last 12 months provides substantial protection to an individual if they become infected with the Iowa strain again. For simulation results not in the low dose range (more than 1 oocyst), the probability of infection for the Iowa strain is approximately 50% to 70% less for those with previous infection in the last 12 months. This could be a substantial planning consideration for a community that has faced significant exposure to *C. parvum* in the last year, bearing in mind that this only considers the Iowa strain.

The results for the highest rate of inactivation (99.96%) always produced a dose in the low dose range of 1 oocyst or less for all combinations of source water oocyst concentrations and water consumption rates. The importance of an effective disinfection process is very evident in the protection of water consumers from *C. parvum* oocysts. The absence of experimental data in the low dose range required the incorporation of human knowledge in that dose range. This model design characteristic has affected the results in the low dose range. Consider that the probability of infection is usually substantially lower for those with a previous Iowa strain infection in the last 12 months as compared to the probability of infection for the population with no previous infection. The probability of infection is approximately equal between the Iowa strain results with previous infection and no previous infection in the low dose range. For example, a dose of 0.03 oocysts was produced for Canadian children at an oocyst concentration of 119 oocysts per L for the source water and disinfection effectiveness of 99.96%, yielding a daily probability of infection of 0.040 for those with no previous infection to the Iowa strain in the last 12 months and 0.038 for those with a previous infection in the last 12 months. This highlights the need for experimental dose response research in the low dose range for *C. parvum*.

A variety of scenarios were considered for the risk characterization including a range of source water oocyst concentrations, several strains of *C. parvum*, different water consumer groups, and several rates of inactivation. In the low dose range the resulting probabilities of infection are higher than the comparable exponential dose response model, reflecting the design decision to be conservative in the low dose range in the absence of experimental data for *C. parvum* in the low dose range. A conservative low dose response region given the dearth of experimental data provides a greater level of vigilance for the health and wellness of a population. However, this may result in significant over estimates of the probability of infection for low doses of *C. parvum* when using these neural network models for risk characterization. Practitioners should carefully consider the assumptions of all dose response models in the low dose region when estimating the level of risk and determining an appropriate set of actions to take. Risk characterizations have been successfully performed with reasonable results using the developed neural network models.

8.0 GENERAL RISK ASSESSMENT LIMITATIONS

The capability of exposure assessment and dose response models to predict risk is limited by the uncertainty and variability of the underlying datasets used to develop the models and the uncertainty introduced by the design of the model itself.

Sources of uncertainty include parameter uncertainty (measurement errors, random errors and systematic bias) and model uncertainty (excluded variables and unconsidered scenarios). For measurement error, very precise measurements are hindered by physical limitations. For example existing methods for determining the concentration of oocysts in a water sample do not provide exact numbers and may not distinguish viable and inactivated oocysts. The use of small datasets can cause random errors that lead to parameter uncertainty. For this thesis a small dataset was used to develop the *C. parvum* dose response model. The high cost of performing dose response challenge experiments means that relatively few subjects are used during an experiment and a small dataset is produced. Systematic errors are reproducible inaccuracies that consistently bias the results in the same direction due to an intrinsic flaw in the experiment or data gathering process. A systematic error that is relevant to predicting the probability of infection for a waterborne pathogen is the lack of consideration for sensitive subpopulations and focusing on healthy adult males that are not representative of the entire population. In this thesis AIDS status was considered in the dose response model, and the drinking water consumption rates for various groups were also incorporated. However, a number of subpopulations were not considered in the dose response and exposure assessment models, such as farm workers who have greater exposure to animal fecal matter.

The developed neural networks used during the exposure and dose response assessments may introduce model uncertainty. For instance, the neural network models may exclude relevant variables or not consider rare sets of conditions that may arise. However all types of models face these problems, whether it be a statistical model, mathematical model or neural network. If a model does not consider a relevant variable or catastrophic scenario then the estimated risk might be grossly in error. Each model presented in this thesis has been carefully developed given the available data and previous studies in the literature. However it is very likely that every possible situation has not been considered and potentially relevant inputs have been missed. This is the nature of modeling, a simplified view of a complex system, thus much detail must be left out.

For risk assessment, variability refers to the variability of input parameters to the dose response and exposure assessment models. For example, water consumption rates vary significantly between age groups and the sensitivity of the AIDS population to *C. parvum* is much greater than healthy adults. It is relatively difficult to include the variation in sensitivity to *C. parvum* across the population due to the ethical constraints of experimenting on sensitive populations and the high cost of performing the dose response challenge experiments. Conversely, it is much easier to determine the approximate differences in water consumption between different groups of consumers. A major limitation of the models presented in this thesis is the ability to incorporate a significant range of variability and granularity, due to dataset constraints. Consider the chlorine dioxide disinfection neural network model; its range for pH is 6 – 11. Now this range might represent the effective range needed for the model to be useful, or there might not have been data available outside of this range.

Finally, the performance of a risk assessment requires that a number of justified assumptions and model design decisions be made. The accuracy of these assumptions directly affects the quality of the risk predictions produced by the risk assessment. A number of assumptions and design decisions were made throughout this thesis, and it is the responsibility of those who use the results of this research to judge the appropriateness of these assumptions and design decisions.

9.0 EXTENSIONS AND CONCLUSIONS

In this thesis three major objectives were achieved: the development of an intelligent dose response model for *C. parvum* that considers three different strains, AIDS status, and previous infection; the creation of a suite of neural network models for performing an exposure assessment of waterborne *C. parvum*; and finally the performance of a risk characterization on waterborne *C. parvum* that considers a variety of scenarios using the developed models.

This was the first known intelligent dose response model developed for *C. parvum*. It outperformed the competing exponential dose response model, but not exceptionally given the added the complexity of the neural network. However, the developed neural network dose response model does have the ability to predict the frequency of infection for a water sample containing combinations of three strains of *C. parvum*. This ability could be useful to risk assessors when the exact strain of *C. parvum* exposure is unknown.

Ozone and chlorine dioxide neural network disinfection models were developed using the SLF learning algorithm, which yielded optimal network structures of three hidden units for the ozone disinfectant and four hidden units for chlorine dioxide. The neural network disinfection models performed better than the temperature corrected Chick-Watson models. The initial and final disinfectant residuals were included as inputs for the neural network disinfection models to simulate the decay of the disinfectant over time. The inclusion of the final residual was determined to not be as important for stable disinfectants, such as chlorine. The relative importance of the model inputs varied between the two disinfection models, but temperature and contact time were the most relevant for both which is in contrast to the classical Chick-Watson model that emphasizes the residual concentration and contact time.

Consumption of cold tap water was used to determine the exposure of an individual to waterborne *C. parvum* because it was assumed that hot water would be stored for a sufficient amount time at a prescribed temperature to inactivate the oocysts. For Canadian water consumption, seasonal changes were found to be irrelevant while age, region and gender were determined to be relatively important. The seasonal neural network consumption model was rejected from use in the risk characterization based on these findings, and the regional neural network consumption model was adopted for use. The regional neural network (RNN) consumption model considers both total tap water consumption and cold tap water consumption to provide flexibility to the risk assessor during an exposure assessment.

Finally, a risk characterization using the various neural network models was performed that considered a variety of oocyst concentrations, three *C. parvum* strains, different rates of inactivation and two water consumer groups. It was noted that experimental studies should be performed in the low dose range to increase the accuracy of models in this region.

9.1 Extensions

Throughout this thesis point estimates have been used for the exposure assessment and the estimates of risk. These point estimates do not provide a span of values that represent the uncertainty and variability of the model input parameters. The use of interval estimates provides a probability distribution or at least a region of confidence, which gives a sense of the uncertainty and variability for the input parameters. Knowledge of input parameter uncertainty and variability is very useful to a risk assessor. It would be worthwhile for researchers to investigate the use of interval estimates for risk assessments using intelligent modeling techniques.

In this thesis only two disinfection processes were presented, ozone and chorine dioxide. A number of other disinfectants and physical removal processes were investigated for potential modeling, including: chlorine, ultra violet light, rapid gravity filtration, inline filtration, and coagulation / flocculation / sedimentation. However, insufficient quality data was available in the literature to develop reasonable models. In the future as more raw data becomes available it would be interesting to create a suite of intelligent models for *C. parvum* disinfection and physical removal processes. This would allow the simulation of complete drinking water treatment processes for *C. parvum* using intelligent models, which would be useful in the design of drinking water treatment plants and the assessment of risk for existing treatment plants.

There are a number of intelligent modeling techniques available to researchers to extend the capabilities of the presented models and achieve greater performance. For example Lau and Musilek (2007) used immune programming to expand on the neural network inactivation models of Janes and Musilek (2007) to produce results that are more easily interpreted than the sometimes-cryptic connection weights of a neural network.

BIBLIOGRAPHY

Abrahart, R.J., See, L. and Kneale, P.E. 2001. Investigating the role of saliency analysis with a neural network rainfall-runoff model. Computers and Geosciences, **27**: 921-928.

Aragon, T., Novotny, S., Enanoria, W., Vugia, D., Khalakdina, A., and Katz, M. 2003. Endemic cryptosporidiosis and exposure to municipal tap water in persons with acquired immunodeficiency syndrome (AIDS): A case control study. BMC Public Health, **3**(2).

Atherholt, T., LeChevallier, M., Norton, W. and Rosen, J. 1998. Effect of rainfall on *giardia* and *Crypto*. Journal of the American Water Works Association. 1998, **90**(9): 66-80.

Baxter, C.W., Stanley, S.J., Zhang, Q. and Smith, D.W. 2002. Developing artificial neural network models of water treatment processes: a guide for utilities. Journal of Environmental Engineering and Science, **1**(3): 201-211.

Beaglehole, R., Bonita, R. and Kjellstrom, T. 1993. Basic Epidemiology. Orient Longman, India.

Brion, G., Neelakantan, T., and Lingireddy, S. 2001. Using neural networks to predict peak *Cryptosporidium* concentrations. Journal of the American Water Works Association, **93**(1): 99-105.

Canada Safety Council, 2005. Heated debate about hot water [online]. Available from http://www.safety-council.org/info/home/hotwater.html [cited 8 January 2007].

Canadian Environmental Health Directorate (CEHD). 1981. Tap water consumption in Canada. Health Protection Branch, Department of the Minister of National Health and Welfare, Ottawa, Canada.

Chappell C., Okhuysen P., Sterling C., Wang C., Jakubowski W., and DuPont H. 1999. Infectivity of *Cryptosporidium parvum* in healthy adults with preexisting anti-*C. parvum* serum immunoglobulin G. American Journal of Tropical Medicine and Hygiene, **60**(1): 157-164.

Chick, H. 1908. An investigation of the laws of disinfection. Journal of Hygiene, **8**: 92-158.

Dawson, R. 2004. Minds and machines: connectionism and psychological modeling. Blackwell Publishing, Oxford, United Kingdom.

Donnelly, J.K., and Stentiford, E.I. 1997. The cryptosporidium problem in water and food supplies. Lebensm.-Wiss. u. –Technol., **30**: 111-120.

Drinking Water Inspectorate (DWI). 1996. Tap water consumption in England and Wales: findings from the 1995 national survey - report number DW10771. Department of Environment, Food and Rural Affairs, United Kingdom.

Duncanson, M., Russell, N., Weinstein, P., Baker, M., Skelly, C., Hearden, M. and Woodward, A. 2000. Rates of notified cryptosporidiosis and quality of drinking water supplies in Aotearoa, New Zealand. Water Research, **34**(15): 3804-3812.

Dupont, H.L, Chappell, C.L., Sterling, C.R., Okhuysen, P.C., Rose, J.B. and Jakubowski, W. 1995. The infectivity of *Cryptosporidium parvum* in healthy volunteers. New England Journal of Medicine, **332**(13): 855-859.

Ershow, A. and Cantor, K. 1989. Total water and tapwater intake in the United States: population based estimates of quantities and sources. Life Sciences Research Office, Federation of American Societies for Experimental Biology, Bethesda, MD.

Fayer, R., Morgan, U. and Upton, S.J. 2000. Epidemiology of cryptosporidium: transmission, detection and identification. International Journal of Parasitology, **30**: 1305-1322.

Gibson, C.J., Haas, C.N., and Rose, J.B. 1998. Risk assessment of waterborne protozoa: current status and future trends. Parasitology, **117**: S205-S212.

Gyurek, L.L, Li, H., Belosevic, M., and Finch, G.R. 1999. Ozone inactivation kinetics of *Cryptosporidium* in phosphate buffer. Journal of Environmental Engineering, **125**(10): 913-924.

Haas, C.N. 2004. Neural networks provide superior description of *giardia lambia* inactivation by free chlorine. Water Research, **30**: 3449-3457.

Haas, C.N., and Rose, J.B. 1996. Distribution of *Cryptosporidium* oocysts in water supply. Water Research, **30**(10): 2251-2254.

Haas, C.N., Rose, J.B., and Gerba, C. 1999. Quantitative Microbial Risk Assessment. John Wiley and Sons, New York, New York.

Haas, C.N., Rose, J.B., Gerba, C., and Regli, S. 1993. Risk assessment of virus in drinking water. Risk Analysis, **13**(5): 545-552.

Heck, S.L., Ellis, G.W., and Hoermann, V. 2001. Modeling the effectiveness of ozone as a water disinfectant using an artificial neural network. Environmental Engineering Science, **18**(3): 205-212.

Hrudey, S.E. and Hrudey, E.J. 2004. Safe Drinking Water: Lessons from Recent Outbreaks in Affluent Nations. IWA Publishing, London, United Kingdom.

Ishikawa, M. 1996. Structural learning with forgetting. Neural Networks, **9**(3): 509-521.

Janes, K.R., and Musilek, P. 2007. Modeling the disinfection of waterborne bacteria using neural networks. Environmental Engineering Science, **24**(4): 448-459.

Janes, K.R., and Musilek, P. In press. Neural network models of *Cryptosporidium parvum* inactivation by chlorine dioxide and ozone. Journal of Environmental Engineering and Science.

Jenkins, M.B., Anguish, L.J., Bowman, D.D., Walker, M.J., and Ghiorse W.C. 1997. Assessment of a dye permeability assay for determination of inactivation *C. parvum* oocysts. Applied and Environmental Microbiology, **63**(10): 2844-3850.

Joseph N., S. Eisenberg, Lei, X., Hubbard, A., Brookhart, M., and Colford, J. 2005. The role of disease transmission and conferred immunity in outbreaks: analysis of the 1993: cryptosporidium outbreak in Milwaukee Wisconsin. American Journal of Epidemiology, **161**(1): 62-72.

Karayiannis, N.B., and Venetsanopoulos, A.N. 1993. Artificial neural networks learning algorithms, performance evaluations, and applications. Kluwer Academic Publishers, Norwell, M.A.

Lau, A. and Musilek, P. Submitted April 2007. Immune programming models of *Cryptosporidium parvum* inactivation by ozone and chlorine dioxide. Information Sciences.

Li, H., Finch, G.R., Smith, D.W., and Belosevic, M. 2001a. Chlorine dioxide inactivation of *Cryptosporidium parvum* in oxidant demand-free phosphate buffer. Journal of Environmental Engineering, **127**(7): 594-603.

Li, H., Finch, G.R., Smith, D.W., and Belosevic, M. 2001b. Sequential inactivation of *Cryptosporidium parvum* using ozone and chlorine. Water Research, **38**(18): 4339-4348.

Li, H., Gyurek, L.L., Finch, G.R., Smith, D.W., and Belosevic, M. 2001c. Effect of temperature on ozone inactivation of *Cryptosporidium parvum* in oxidant demand-free phosphate buffer. Journal of Environmental Engineering, **127**(5): 456-467.

MacKenzie, W.R., Hoxie, N.J., Proctor, M.E., Gradus, M.S., Blair, K.A., Peterson, D.E., Addiss, D.G., Fox, K.R., Rose, J.B., and Davis, J.P. 1994. A massive outbreak in Milwaukee of *Cryptosporidium* infection transmitted through the public water supply. New England Journal of Medicine, **331**: 161-167.

Makri, A., Modarres, R., and Parkin, R. 2004. Cryptosporidiosis susceptibility and risk: a case study. Risk Analysis, **24**(1): 209-220.

Martins, C.A.P, and Guerrant, R.L. 1995. Cryptosporidium and cryptosporidiosis. Parasitology Today, **11**(11): 434-436.

Messner, M.J, Chappell, C.L., and Okhuysen, P.C. 2001. Risk assessment for cryptosporidium: a hierarchical bayesian analysis of human dose response data. Water Research, **35**(16): 3934-3940.

Office of Environmental Health Hazard Assessment (OEHHA). 2000. Technical support document for exposure assessment and stochastic analysis. OEHHA, Sacramento, California.

Okhuysen, P., Rich, S., Chappell, C., Grimes, K., Widmer, G., Feng, X., and Tzipori, S. 2002. Infectivity of a *Cryptosporidium parvum* isolate of corvine origin for healthy adults and interferon-γ knockout mice. Journal of Infectious Diseases. **185**: 1320-1325.

Pouillot, R., Beaudeau, P., Denis, J.B., and Derouin, F. 2004. A quantitative risk assessment of waterborne cryptosporidiosis in France using second order monte carlo simulation. Risk Analysis, **23**(1): 1-17.

Rognvaldsson, T. S. 1998. A simple trick for estimating the weight decay parameter. *In* Neural networks: tricks of the trade. *Edited by* G.B. Orr and K. Muller. Springer, USA, pp. 71–93.

Rose, J.B., Huffman, D.E., and Gennaccaro, A. 2002. Risk and control of waterborne cryptosporidiosis. Federation of European Microbiological Societies Microbiology Reviews, **26**: 113 – 123.

Roseberry, A. and Burmaster, D. 1992. Lognormal distributions for water intake by children and adults. Risk Analysis, **12**: 99-104.

Silvert, W., and Baptist, M. 2000. Can neuronal networks be used in data-poor situations? *In* Artificial neuronal networks: application to ecology and evolution. *Edited by* S. Lek and J.F. Guegan. Springer-Verlag, Berlin, pp. 241-248.

Teunis, P.F., Chappell, C.L., and Okhuysen, P.C. 2002. Cryptosporidium dose response studies: variation between hosts. Risk Analysis, **22**(3): 475 -485.

Teunis P.F.M., Medema, G.J., Kruidenier, L., and Havelaar, A.H. 1997. Assessment of the risk of infection by cryptosporidium or giardia in drinking water from a surface water source. Water Research, **31**(6): 1333 -346.

Thomas, S.P and Hrudey, S.E. 1997. Risk of Death in Canada: What We Know and How We Know It. University of Alberta Press, Edmonton, Alberta.

Watson, H.E. 1908. A note on the variation of the rate of disinfection with change in the concentration of the disinfectant. Journal of Hygiene, **8**: 536-542.

World Health Organization (WHO). 2004. Guidelines for drinking water quality: third edition. WHO Library, Geneva.

Yang. 2003. A neural network model for dose-response of foodborne pathogens. Applied Soft Computing, **3**(2): 85–96.

Yu C. 2004. Soft computing approaches for microbial food safety approaches. University of Guelph, Guelph, Ontario.

Zhang, Q., Yang, S.X., Mittal, G.S., and Yi, S.J. 2002. Prediction of performance indices and optimal parameters of rough rice drying with neural networks. Biosystems Engineering, **83**(3): 281–290.

APPENDIX A: RISK CHARACTERIZATION RESULTS

Ozone Effectiveness (Rate of Inactivation)	Cold Tap Water Consumption (L per day)	Dose (Oocysts)	Strain of *C. parvum*	AIDS Status (X = positive)	Previous Infection in the last year (X=previous infection)	Probability of Infection
46.29%	0.64	4.12	Iowa			0.117
			UCP			0.111
			TAMU			0.255
			Iowa		X	0.059
			Iowa	X		0.519
	0.97	6.25	Iowa			0.139
			UCP			0.127
			TAMU			0.386
			Iowa		X	0.064
			Iowa	X		0.605
96.45%	0.64	0.27	Iowa			0.054
			UCP			0.054
			TAMU			0.036
			Iowa		X	0.043
			Iowa	X		0.158
	0.97	0.41	Iowa			0.059
			UCP			0.059
			TAMU			0.043
			Iowa		X	0.044
			Iowa	X		0.187
99.96%	0.64	0.0031	Iowa			0.034
			UCP			0.028
			TAMU			0.016
			Iowa		X	0.036
			Iowa	X		0.054
	0.97	0.0047	Iowa			0.035
			UCP			0.029
			TAMU			0.016
			Iowa		X	0.036
			Iowa	X		0.057

Table A.1: Simulation #1 – 12 oocysts/L for source waters

Ozone Effectiveness (Rate of Inactivation)	Cold Tap Water Consumption (L per day)	Dose (Oocysts)	Strain of *C. parvum*	AIDS Status (X = positive)	Previous Infection in the last year (X=previous infection)	Probability of Infection
46.29%	0.64	40.91	Iowa			0.338
			UCP			0.257
			TAMU			0.956
			Iowa		X	0.105
			Iowa	X		0.899
	0.97	62.00	Iowa			0.413
			UCP			0.302
			TAMU			0.980
			Iowa		X	0.122
			Iowa	X		0.930
96.45%	0.64	2.70	Iowa			0.100
			UCP			0.097
			TAMU			0.168
			Iowa		X	0.055
			Iowa	X		0.436
	0.97	4.09	Iowa			0.117
			UCP			0.111
			TAMU			0.253
			Iowa		X	0.059
			Iowa	X		0.518
99.96%	0.64	0.03	Iowa			0.040
			UCP			0.036
			TAMU			0.020
			Iowa		X	0.038
			Iowa	X		0.080
	0.97	0.046	Iowa			0.042
			UCP			0.039
			TAMU			0.022
			Iowa		X	0.039
			Iowa	X		0.088

Table A.2: Simulation #2 – 119 oocysts/L for source waters

Ozone Effectiveness (Rate of Inactivation)	Cold Tap Water Consumption (L per day)	Dose (Oocysts)	Strain of *C. parvum*	AIDS Status (X = positive)	Previous Infection in the last year (X=previous infection)	Probability of Infection
46.29%	0.64	85.94	Iowa			0.478
			UCP			0.343
			TAMU			0.989
			Iowa		X	0.138
			Iowa	X		0.948
	0.97	130.25	Iowa			0.566
			UCP			0.401
			TAMU			0.995
			Iowa		X	0.163
			Iowa	X		0.964
96.45%	0.64	5.68	Iowa			0.133
			UCP			0.123
			TAMU			0.351
			Iowa		X	0.063
			Iowa	X		0.585
	0.97	8.61	Iowa			0.159
			UCP			0.142
			TAMU			0.514
			Iowa		X	0.068
			Iowa	X		0.670
99.96%	0.64	0.064	Iowa			0.043
			UCP			0.041
			TAMU			0.023
			Iowa		X	0.039
			Iowa	X		0.096
	0.97	0.097	Iowa			0.046
			UCP			0.044
			TAMU			0.026
			Iowa		X	0.040
			Iowa	X		0.109

Table A.3: Simulation #3 – 250 oocysts/L for source waters