Computers, Environment and Urban Systems 34 (2010) 189–203

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Use of genetic algorithm and neural network approaches for risk factor selection: A case study of West Nile virus dynamics in an urban environment

Debarchana Ghosh [a,*], Rajarshi Guha [b]

[a] Department of Geography, Kent State University, 413 McGilvrey Hall, Kent, OH 44240, USA
[b] National Institute of Health Chemical Genomics Center, 9800 Medical Center Drive, Rockville, MD 20852, USA

ARTICLE INFO

Article history:
Received 30 September 2009
Received in revised form 22 February 2010
Accepted 23 February 2010

Keywords:
West Nile virus
Risk factors
Genetic algorithm
Variable selection
Neural network

ABSTRACT

The West Nile virus (WNV) is an infectious disease spreading rapidly throughout the United States, causing illness among thousands of birds, animals, and humans. Yet, we only have a rudimentary understanding of how the mosquito-borne virus operates in complex avian–human environmental systems coupled with risk factors. The large array of multidimensional risk factors underlying WNV incidences is environmental, built-environment, socioeconomic, and existing mosquito abatement policies. Therefore it is essential to identify an optimal number of risk factors whose management would result in effective disease prevention and containment. Previous models built to select important risk factors assumed a priori that there is a linear relationship between these risk factors and disease incidences. However, it is difficult for linear models to incorporate the complexity of the WNV transmission network and hence identify an optimal number of risk factors objectively.

There are two objectives of this paper, first, use combination of genetic algorithm (GA) and computational neural network (CNN) approaches to build a model incorporating the non-linearity between incidences and hypothesized risk factors. Here GA is used for risk factor (variable) selection and CNN for model building mainly because of their ability to capture complex relationships with higher accuracy than linear models. The second objective is to propose a method to measure the relative importance of the selected risk factors included in the model. The study is situated in the metropolitan area of Minnesota, which had experienced significant outbreaks from 2002 till present.

1. Introduction

West Nile virus (WNV), first isolated in Uganda in 1937, is a vector-borne infectious disease of global public health concern. The virus is transmitted to humans and other mammals by infected mosquitoes that acquire the virus by feeding on WNV-infected birds (CDC., 1999). During previous outbreaks, the infection in humans was not considered fatal, and infections typically varied from asymptomatic symptoms to mild illness with fever, rash, and headache. However, during the recent outbreaks in southern Romania (1999), the Volga delta of Russia (1999) and the Northeastern United States (between 1999 and 2000) there were instances of the more severe form of illness, i.e., West Nile encephalitis (inflammation of the brain), and meningitis (inflammation of the lining of the brain and spinal cord), both of which can be fatal. In the western hemisphere, the United States has the highest number of infected cases. Since its initial epicenter in New York in 1999, the virus has spread rapidly north, south, and west, causing seasonal epidemics and illness among thousands of birds, mosquitoes, humans, and horses.

It is challenging to identify important risk factors of WNV because the virus propagates via complex interrelationships between human, avian, and mosquito habitat systems. The multi-dimensional risk factors underlying WNV incidences are environmental (temperature, precipitation, vegetation, hydrologic features and parks), socioeconomic (occupation, income, housing age and condition), built-environment (catch basins, construction sites, ditches, scrap-tire stockpiles and sewers), and existing mosquito abatement policies. Previous models built to identify important risk factors assumed *a priori* that there is a linear relationship between these risk factors and WNV incidences (Bowman, Gumel, Driessche, Wu, & Zhu, 2005; Brownstein et al., 2002; Cooke, Katarzyna, & Wallis, 2006; David, Mak, MacDougall, & Fyfe, 2007; Diuk-Wasser, Brown, Andreadis, & Fish, 2006; Gibbs et al., 2006; Lian, Warner, Alexander, & Dixon, 2007; Ruiz, Tedesco, McTighe, Austin, & Kitron, 2004). However, it is difficult for linear models to incorporate the complexities of the WNV transmission network. As a result, previous investigations were not able to rigorously identify the *optimal* risk factors. It is only by simultaneously examining

* Corresponding author. Tel.: +1 3306723220; fax: +1 3306724304.
E-mail addresses: dghosh@kent.edu (D. Ghosh), guhar@mail.nih.gov (R. Guha).

0198-9715/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2010.02.007

all of the factors as well as their interplay that we can hope to risk factors underlying WNV transmission network.

There are two objectives of this paper. The first is to use a combination of genetic algorithm (GA) and computational neural network (CNN) approaches to build a model incorporating the *nonlinearity* between WNV disease incidences and hypothesized risk factors. Here GA is used for risk factor (variable) selection and CNN for model building. There are several advantages of using neural network algorithms in understanding a complex health outcome over the linear and traditional nonlinear models (logistic and Poisson). These are as follows: (1) their ability to capture complex relationships between risk factors and disease occurrence, (2) their good predictive capability, (3) that they do not require an *a priori* distribution to be specified because they learn the relationships between the input and the output from the dataset itself and (4) the lack of rigid assumptions of normality and homoscedasticity. The second objective is to propose a method to measure the relative importance of the selected risk factors included in the model. The interpretation technique, called '*broad*', is essentially a sensitivity analysis of the neural network, which ranks the risk factors in order of their significant contribution to the predictive capability of the CNN model.

The study is situated in the Twin Cities Metropolitan Area (TCMA) of Minnesota, United States (Fig. 1). The virus first reached Minnesota in 2002, creating epidemiological 'hotspots' in the metropolitan area in 2003, 2006, and 2007. In 2003, clusters of WNV incidences were found in the cities of Minneapolis and Saint Paul with 26 human cases, 285 infected dead birds, and approximately 1400 infected mosquito pools. The number of infected dead birds (479) increased dramatically in 2006, and they were mostly clustered around the twin cities of Minneapolis and Saint Paul. Fig. 2 further exhibits the urban-centric nature of WNV transmission in Minnesota. This pattern was similar to other WNV outbreaks observed in the urban areas of northeastern United States (Brown, Childs, Diuk-Wasser, & Fish, 2008; Hayes et al., 2005; Mostashari, Martin, Hartman, Miller, & Kulasekera, 2003), Chicago, Detroit (Ruiz, Walker, Foster, Haramis, & Kitron, 2007), and Georgia (Gibbs et al., 2006).

The modeling techniques employed to predict and identify the important risk factors of WNV infection can be broadly divided into three groups – linear, nonlinear, and spatial. The ordinary least squares (OLS) method is the common statistical technique used in the linear category (Ezenwa et al., 2007; Mongoh, Khaitsa, & Dyer, 2007; Pradier, Leblond, & Durand, 2008; Yiannakoulias, Schopflocher, & Svenson, 2006). Even though linear models are simple to interpret, they suffer from several disadvantages, especially when dealing with complex health outcomes. In reality, the transmission of a multi-host infection such as WNV is nonlinear, and therefore it is challenging for linear models to account for the nonlinearity of the relationships between disease outcomes and risk factors. Moreover, the inflexibility of OLS models due to their rigid assumptions (normality, independent observations, and homoscedasticity) makes them inappropriate to model complex phenomena (Mennis & Diansheng, 2009).

The nonlinear models employed in WNV research can be divided into two broad groups, intrinsically linear and intrinsically nonlinear models. The former class of models can be transformed into a *linear* function and subsequently analyzed using linear models, and thus are not *entirely* nonlinear. Examples of these types of methods include the Poisson, logit, and probit regression models. However, in the intrinsically nonlinear models, the nonlinear form of the model cannot be transformed to a linear form. Examples of this type of models include the general growth model, maximum likelihood, and neural network algorithms. Previous research also

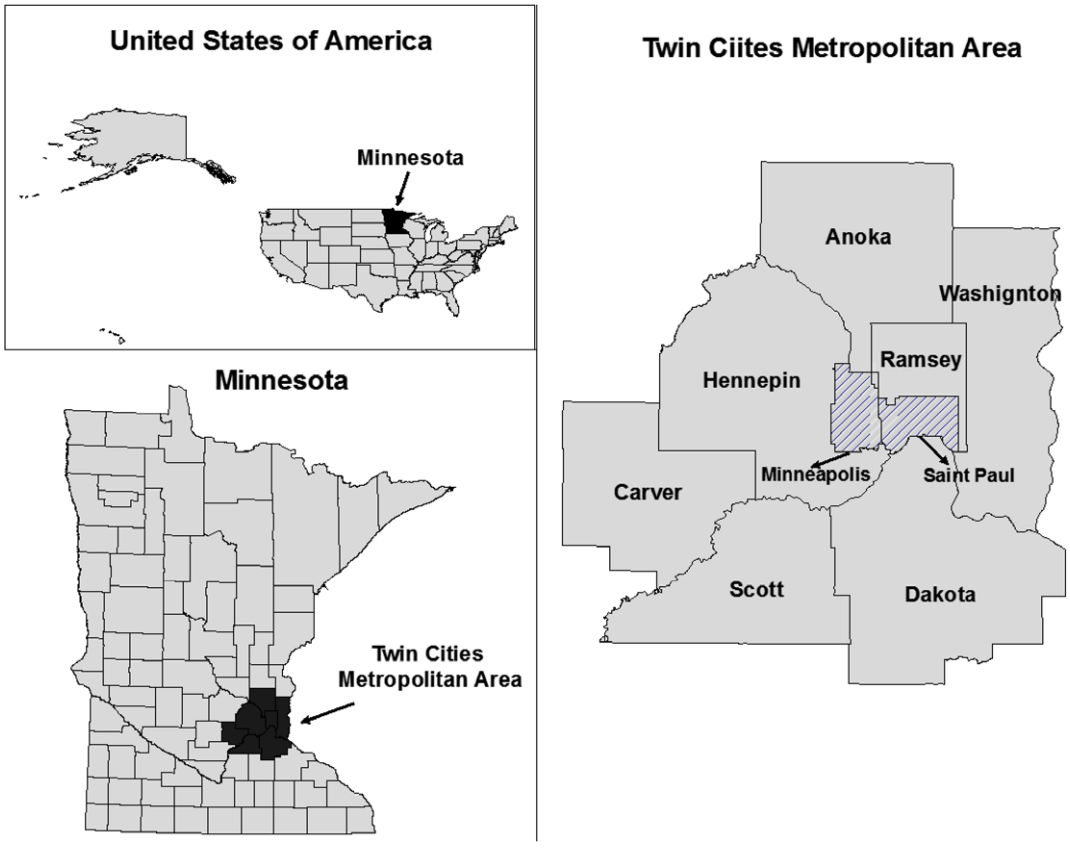

Fig. 1. Location of the twin cities metropolitan area of Minnesota, United States.

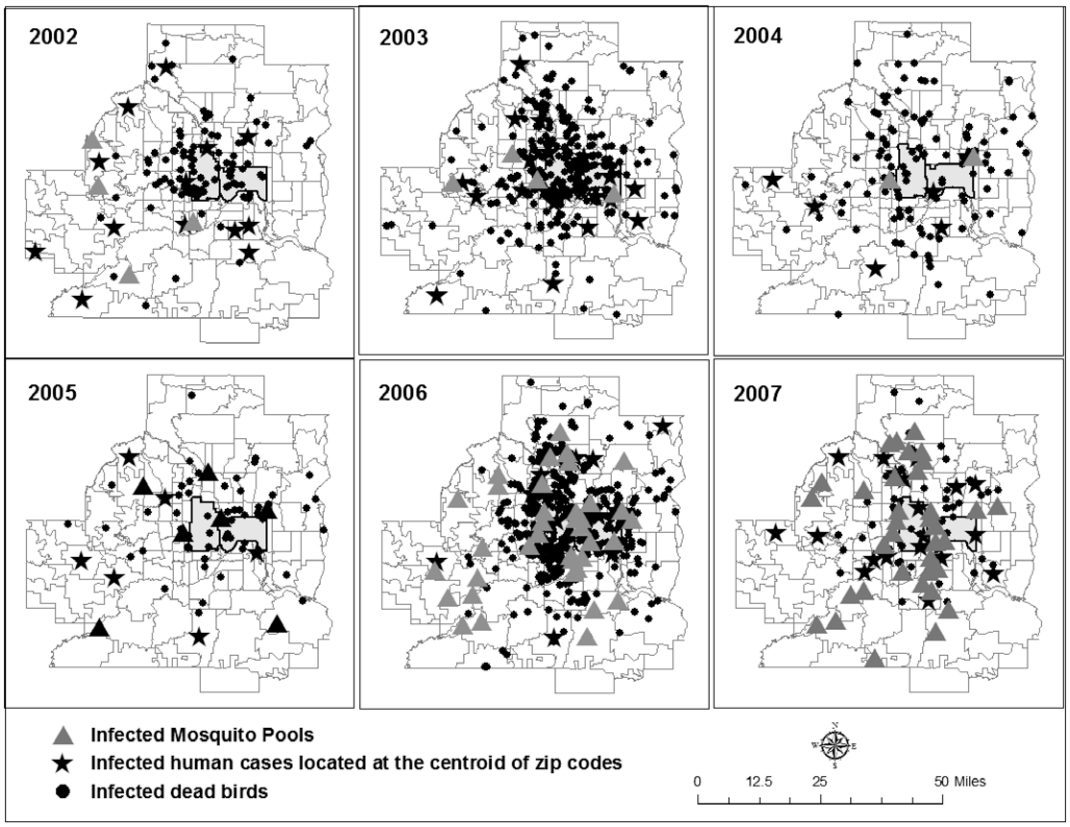

Fig. 2. Spatial overlay of West Nile virus incidences showing the urban-centric nature of the transmission of the virus in the metropolitan area of Minnesota.

shows examples of modeling WNV disease occurrence by logistic (Brown et al., 2008; Diuk-Wasser et al., 2006; Gibbs et al., 2006; LaBeaud et al., 2008; Leblond, Sandoz, Lefebvre, Zeller, & Bicout, 2007; Murray et al., 2006; Ruiz et al., 2004; Shaman, Day, & Stieglitz, 2005) and Poisson regressions (Roberts & Foppa, 2006; Yiannakoulias et al., 2006). There is no doubt that these studies contributed significantly to the limited knowledge of a rapidly spreading virus, but can be improved further by using sophisticated *nonlinear* models. For both logistic and Poisson regression models, *a priori* functional form or a statistical distribution between the response variable (WNV infection) and the predictors (risk factors) has to be specified. However, given the novelty and complexities of the transmission characteristics of WNV in the western world, the interaction of avian and mosquito habitats, and the catalytic effect of environmental, built-environment, and socio-demographic risk factors, the assumption of a predefined distribution is simplistic. Therefore, it is difficult for these intrinsically linear models to incorporate the nonlinearity between disease occurrence and the risk factors.

The spatial models used in the WNV research are generalized linear mixed models (GLMM) (Johnson, 2008; Yiannakoulias & Svenson, 2007; Yiannakoulias et al., 2006). The main advantage of the spatial models is their ability to account for spatial autocorrelation, especially when myriad environmental, climatic, and social risk factors are included in the models as predictors. However, this group of model also suffers from similar weaknesses to those of the linear and logistic regression models: first, the assumption of a predefined relationship between response and predictor variables, and second, the lack of nonlinearity. Attempts to capture the nonlinearity between the occurrence of WNV incidences and hypothesized risk factors in the literature are limited.

This paper is organized as follows. Section 2 describes the data and discusses the underlying details of the modeling, optimization, and cross-validation techniques used in this work. Section 3 explains the steps involved in developing a nonlinear WNV analysis model based upon approaches discussed in the previous section. Section 4 interprets the results obtained from the model, and finally, Section 5 concludes the paper.

2. Materials and method

This section describes the data and the underlying details of the modeling and optimization techniques used in this work.

2.1. Data description

The data for the year 2006, with 479 WNV infected dead birds, was used to build the model. The risk factors, as predictor variables, belonged to the four categories of environmental, built-environment, proximity, and existing vector control programs. After checking for collinearity, missing values, and data ranges (variation), 32 potential risk factors were included in the model. Maximum daily temperature, daily precipitation, and land cover variables were grouped in the environmental category. The built-environment factors mainly included density of catch basins, density of ditches, area of parks (open green space), housing density, and housing age. Proximity to features such as lakes, wastewater discharge points, golf courses, trails, shrub swamps, wooded swamp, and bogs was also considered. The variables in the final category of vector control policies were frequency and percentage of public land survey (PLS) units treated for larvicide and adulicide. All the variables were processed and aggregated at the US census zip code level. Tables 1 and 2 show the sources and summary statistics of all the variables included in the model.

Table 1
Descriptive summary of variables in the model.

Variables	Min.	Median	Mean	Max.	Unit	Source
Response variable						
WNV infected dead birds	0	2	3	17	Count	MDH
Predictor variables						
Daily max. temperature	75.39	82.93	79.52	98	F	NCDC, COOP
Daily precipitation	0	0.08804	0.1494	1	Inches	NCDC, COOP
Developed, open space	0	7.647	8.92	24.72	%	NLCD
Developed, low density	0.6151	17.03	18.8	64.28	%	NLCD
Developed, medium density	0	12.52	12.77	39.11	%	NLCD
Developed, high density	0	4.291	9.12	81.24	%	NLCD
Shrub/scrub	0	0.4003	1.054	6.838	%	NLCD
Pasture/hay	0	3.565	9.293	42.94	%	NLCD
Cultivated crops	0	1.31	12.5	76.74	%	NLCD
Density of catch basins (wet)	0	23.18	43.94	383.1	sq. mile	MMCD
Density of catch basins (dry)	0	171.5	200.4	740.1	sq. mile	MMCD
Density of ditches	0	0	0.1238	1.145	sq. mile	MnDOT
Housing density	1	1.73	1.914	5.029	Acre	MC
Age of houses	0	37	41.21	90	Years	MC
Density of bike paths	0	0.4244	0.8954	4.682	sq. mile	MetroGIS
Percentage area of parks	0	6.508	7.827	41.61	%	MetroGIS
Flooded basins and flat (D)	177.3	503.6	602.9	2394	Miles	MMCD
Inland fresh meadow (D)	108.2	440.2	586	2143	Miles	MMCD
Inland shallow marsh (D)	167.2	530.3	792.2	2990	Miles	MMCD
Inland deep fresh marsh (D)	136.2	439.8	555.4	1854	Miles	MMCD
Inland fresh open water (D)	420.7	1067	1603	13,620	Miles	MMCD
Shrub swamps (D)	558.1	4242	5008	17,250	Miles	MMCD
Wooded swamp (D)	1359	8295	8977	38,800	Miles	MMCD
Bog (D)	661.4	5655	6996	25,300	Miles	MMCD
Lake (D)	60.59	440.5	524.3	1872	Miles	MetroGIS
Park (D)	117.7	442.1	1147	9027	Miles	MetroGIS
Water discharge points (D)	240.5	1960	4334	19,930	Miles	MetroGIS
Golf courses (D)	1202	2897	3838	21,490	Miles	TLG
Trail (unpaved) (D)	138.5	2639	5309	35,680	Miles	MetroGIS
Percentage of larvicide treat	1.639	92.86	76.9	100	%	MMCD
Frequency of larvicide app.	2	28	31	83	Count	MMCD
Percent of adulticide treat	0	21.28	26.26	100	%	MMCD
Frequency of adulticide app.	1	3	3	19	Count	MMCD

Note: All the variables are aggregated and processed at the US census zip code level.
*The abbreviations for the data sources are described in Table 2.

Table 2
Data sources.

Acronyms	Full description
MDH	Minnesota Department of Health
COOP	Climatological Observation Stations
NCDC	National Climatic Data Center
NLCD	National Land Cover Data
MMCD	Metropolitan Mosquito Control District, Minnesota
MnDOT	Minnesota Department of Transport
MC	Metropolitan Council
MetroGIS	GIS data house of Minnesota
TLG	The Lawrence Group

2.2. Optimization method – genetic algorithm

In social, environmental, and health sciences, statistical and mathematical modeling begins by considering a large number of predictor variables, especially when investigating complex health outcomes (Lyme disease, WNV, malaria, SARS, etc.), land use /land cover analysis, or other human–environment interactions. The pool of predictor variables often includes multi-dimensional factors such as environmental, built-environment, social, economic, and demographic variables. Even though including myriad variables incorporates issues associated with missing variables, it often leads to *over fitting*. This is especially true when the sample size n is small. Based on the theory of parsimony, and to minimize over fitting, the goal here is to include the optimal or minimum number of predictors that will explain the maximum variation of the response variable. Thus, we must objectively select a *good* or *best* subset of predictor variables.

There exist several statistical methods that perform variable selection such as stepwise regression, backward, and forward selection. However, these methods are generally restricted to linear and logistic regression methods. A more important reason for not considering these methods is due to a number of their drawbacks, including falsely calculated narrow confidence intervals (Altman & Anderson, 1989), incorrect p-values, biased regression coefficients, and problems with multi-collinearity and so on (Guha, Stanton, & Jurs, 2005). The forward and backward selection algorithms, by their nature, will ignore certain combinations of variables because in the former case variables are based on the current subset that has already been selected and in the latter case variables removed from consideration are not considered again. This will have significant effects in modeling a phenomenon where the interrelationships and combinations of predictor variables are important. Therefore, we used an alternative optimization algorithm to conduct variable selection. Among the several deterministic and stochastic optimization methods available, we have used genetic algorithms (GA) to identify a *best* subset of predictor variables.

GA is a stochastic optimization technique based on the concepts of biological evolution, such as reproduction, crossover, and mutation (Shad, Mesgari, Abkar, & Shad, 2009). As a result, much of the terminology used in the description of GA's is adapted from the field of biological evolution. Examples include population,

individual, chromosome, fitness value, crossover, mutation, and child population. These terms are defined in the following paragraphs. When using GA for variable selection, a *chromosome* is defined as a subset of predictor variables (of user specified size), selected from a pool of predictors. An *individual* is comprised of a chromosome and an associated *fitness value.* A *population* is defined as a collection of individuals.

The first step in a GA is to initialize the population by randomly generating a set of individuals. The exact number is user specified and depends on the size of the pool of predictors and the size of predictor subsets desired (Guha, 2005). Each subset is then used to build a model (which in our case study is a CNN model), and the root mean square error (RMSE) for each model is used as the fitness of the individual (Eq. (1)). The population is then ranked based on the fitness value for each individual.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2} \tag{1}$$

The next step is to create a child population, as described below:

1. Individuals are selected to create a mating list, equal in size to the current population.
2. Those individuals with fitness value greater than the population average are automatically placed in the mating list.
3. The remaining positions are filled by using a roulette wheel selection procedure to select individuals from the current population (Guha, 2005)
4. Finally, a child population is generated by randomly selecting pairs of individuals from the mating list and performing genetic operations such as *crossover* and *mutation*.

Crossover involves the swapping of portions of the chromosomes (i.e., the predictor variables) of a pair of individual (Forrest, 1993). The goal of the crossover function is to generate new individuals reflecting good features of their parent population. The second genetic function is mutation and is performed on individuals in the child population. Mutation is performed by randomly changing a part of the chromosome (i.e., randomly replacing one of the predictor variable with another one from the predictor pool) of an individual, and its main function is to maintain the diversity of the population.

With the application of these operations, a child population is generated, whose individuals are then evaluated and ranked based on their fitness values. The second-generation population is then generated by randomly selecting individuals from the top 50% of the child population and randomly selecting an equal number from the remainder. The whole process is repeated for a user specified number of iterations (stopping criterion), typically 100, but this is problem dependent. Finally, the top ranked individuals (i.e., top ranked predictor subsets) with their fitness value (RMSE) are reported to the user. We should mention at this point that the current implementation of the algorithm does not employ a bit string as such. Instead, each individual is simply represented as a vector of predictor indices. All genetic operations operate on this "index vector". Thus, mutation simply replaces one of the indexes with the index of a randomly chosen predictor. The fitness of the individual is then evaluated by the RMSE obtained from a CNN model using the predictors as identified by the individual's index vector (i.e., chromosome). Fig. 3 shows the integration of GA and CNN in this work.

2.3. Computational neural networks

CNN belong to the intrinsically nonlinear modeling category, in which the functional form cannot be transformed to a linear form. A CNN essentially attempts to mimic the behavior of a human brain and thus a fundamental characteristic of these algorithms is the ability to *learn* the relationships present within a dataset. It resembles the behavior of a human brain in the following aspects: (1) the basic unit of operation in a neural network is the *neuron*; (2) knowledge is acquired by the network from the environment (dataset) through a learning process and (3) interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge (Fischer, 1998; Guha & Jurs, 2005).

Haykin (2001) described a neural network as "an extensive parallel distributed processor made up of simple processing units, which has a natural propensity and ability to store knowledge through learning and making it available for further use". There are a large variety of neural network algorithms, namely feed forward, back propagation learning, probabilistic, radial basic function, self-organizing maps, and structural learning and forgetting algorithms (Haykin, 2001). Among these algorithms, we have used the feed-forward neural network algorithm for this particular work, mainly for two reasons. First, feed-forward algorithms are simpler than the other algorithms mentioned above. Second, the technique used to interpret the CNN models in this work has *only* been developed for feed-forward neural network algorithms (Guha & Jurs, 2005).

The structure of a feed-forward neural network model includes three layers, which are fully connected. The first layer is called the input layer, and each neuron in this layer corresponds to a predictor variable in the model. The second layer is called the hidden layer and is responsible for nonlinearly combining the values of the predictor variables. The final layer is called the output layer, whose output is the predicted value of the response variable. The term fully-connected indicates that all the neurons in a given layer are connected to all the neurons in the next layer.

The mechanism of a neural network is based on the neurons. At each layer, a neuron accepts input values and weights associated with the nonlinear functions in the preceding layer, and these values are then transferred to the next layer by a *transfer function.* The main advantage of a neural network is that the transfer function is generally nonlinear in nature. A number of transfer functions are described in the literature (Haykin, 2001), and the implementation used in this study applies a sigmoidal function given by the following formula:

$$O = \frac{1}{1 + \exp(-\sum x_i w_i + b)} \tag{2}$$

where O is the output of the neuron, x_i is the output value of the ith neuron in the preceding layer, w_i is the weight for the connection between this neuron and the ith neuron in the preceding layer, and b is the value of the bias term (Guha, 2005). The weights between the input and the hidden layer neurons allow the network to be configured so that the more important predictor variables will make greater contributions to the hidden layer neurons. The weights and biases, transferring from one layer to another, allow the network to *learn* the features present in the dataset and make accurate predictions.

Thus, the next important step in the neural network algorithm is to obtain an optimal set of weights and biases. Typically the numbers of weights and biases are specified by the structure of the neural network, i.e., by the number of input (predictor variables) and hidden layer neurons. Increasing the number of hidden neurons will improve the prediction, but with a cost of over fitting. One rule of thumb to determine the appropriate number of weights and biases is that the total number of parameters should be less than half the size of the training set used to build the model.

Next, a CNN model is initialized to train the network. To do so, we used the *nnet* function from the package nnet in the *R* statistical

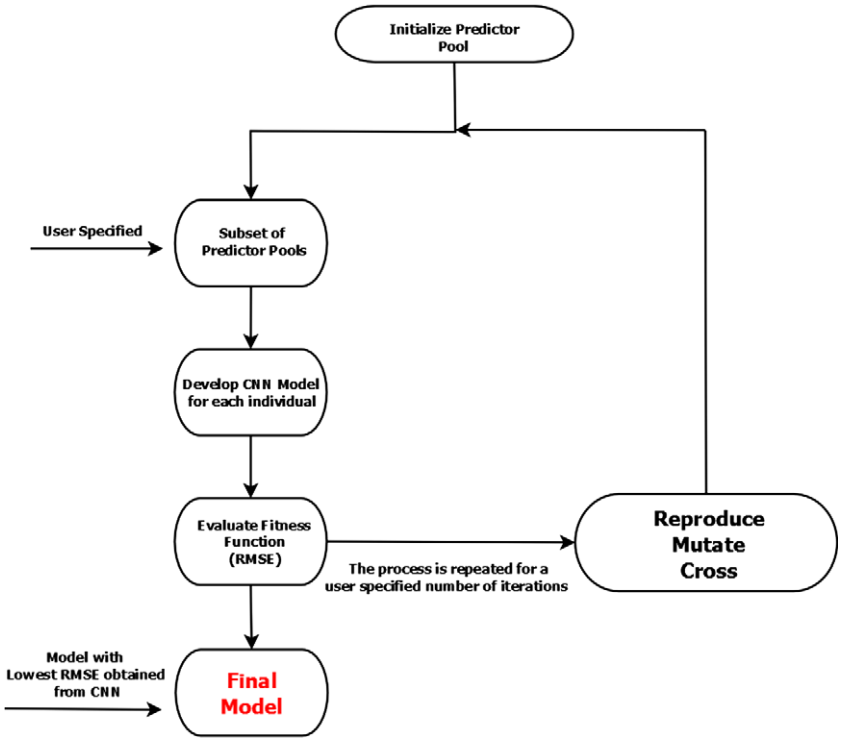

Fig. 3. A flow chart describing the implementation of genetic algorithm and neural network.

software programming language (http://cran.r-project.org/). In the *nnet* function, the network is initialized with a set of weights and biases using a random number generator and the network is trained by the Broyden–Fletcher–Golfarb–Shanno (BFGS) optimization algorithm. An important feature of the training phase is that it is supervised. That is, to train the network, observed values for the training set are required. It is also important to note that since the network is initialized using random numbers, multiple runs of the *nnet* function can result in slightly different sets of optimal weights and biases. As a result, good practice suggests that the final result is obtained using an ensemble of models obtained from multiple runs of the *nnet* function. Once the model has been trained, its quality can be evaluated by noting the R^2 (square of the correlation between observed and predicted y variable) and the RMSE values.

2.4. Interpretation of computation neural networks

The interpretation technique, called '*broad*', is essentially a sensitivity analysis of the neural network, which is viewed here as a measure of predictor importance. The algorithm to measure predictor importance, adapted from the literature of computational chemistry (Guha & Jurs, 2005) are as follows:

1. To start, a neural network model is trained and validated. The RMSE obtained from this model is defined as the base RMSE.
2. The first input predictor variable (say distance to bogs) is randomly scrambled, and then the neural network model is rebuilt and used to predict the number of WNV infected dead birds. Because the values of this predictor are shuffled, one would expect the correlation between distance to bogs and WNV occurrence in birds to be obscure. As a result, the RMSE calculated from these new predictions should be larger than the base RMSE, and the difference between these two RMSEs indicates the importance (contribution) of distance to bogs to the model's predictive ability. That is, if an input factor contributes significantly to the model's predictions, scrambling that factor will lead to a greater loss of predictive power (as measured by the RMSE value) than for an input factor that does not play such an important role in the model.
3. This procedure is then repeated for all the predictor variables present in the model.
4. Finally the predictor variables are ranked in order of their importance (difference between the base RMSE and RMSEs obtained from the new predictions).

2.5. Cross-validation

Once the model is built, cross-validating is the next important step. Model validation can be conducted using several cross-validation techniques, both internally and externally. In this work, the Leave-one-out cross validation or LOO method is used as an internal technique (Golbraikh & Tropsha, 2002). The Q^2 value obtained by using the LOO cross-validation procedure is an alternative to R^2 That is, a neural network model with the same structure (same number of input variables, hidden neurons, and output neuron) is generated using the whole dataset excluding one point. The response value for this point is then predicted using the model, and this procedure is repeated for all the points in the dataset. The R^2 for these predictions is denoted by Q^2. Typically, if the R^2 and Q^2 values are in the range of 10–15%, then the model is considered good, with high predictive ability and generalizability. For external cross-validation methods, the RMSE and the R^2 values obtained from the new datasets are observed. Ideally one would expect higher R^2 and RMSE values, usually within 10% of the original response variable.

3. West Nile virus analysis model

Section 3.1 describes the steps involved in building and selecting the best neural network WNV analysis model for the TCMA. Section 3.2 defines the CNN architecture and statistics of

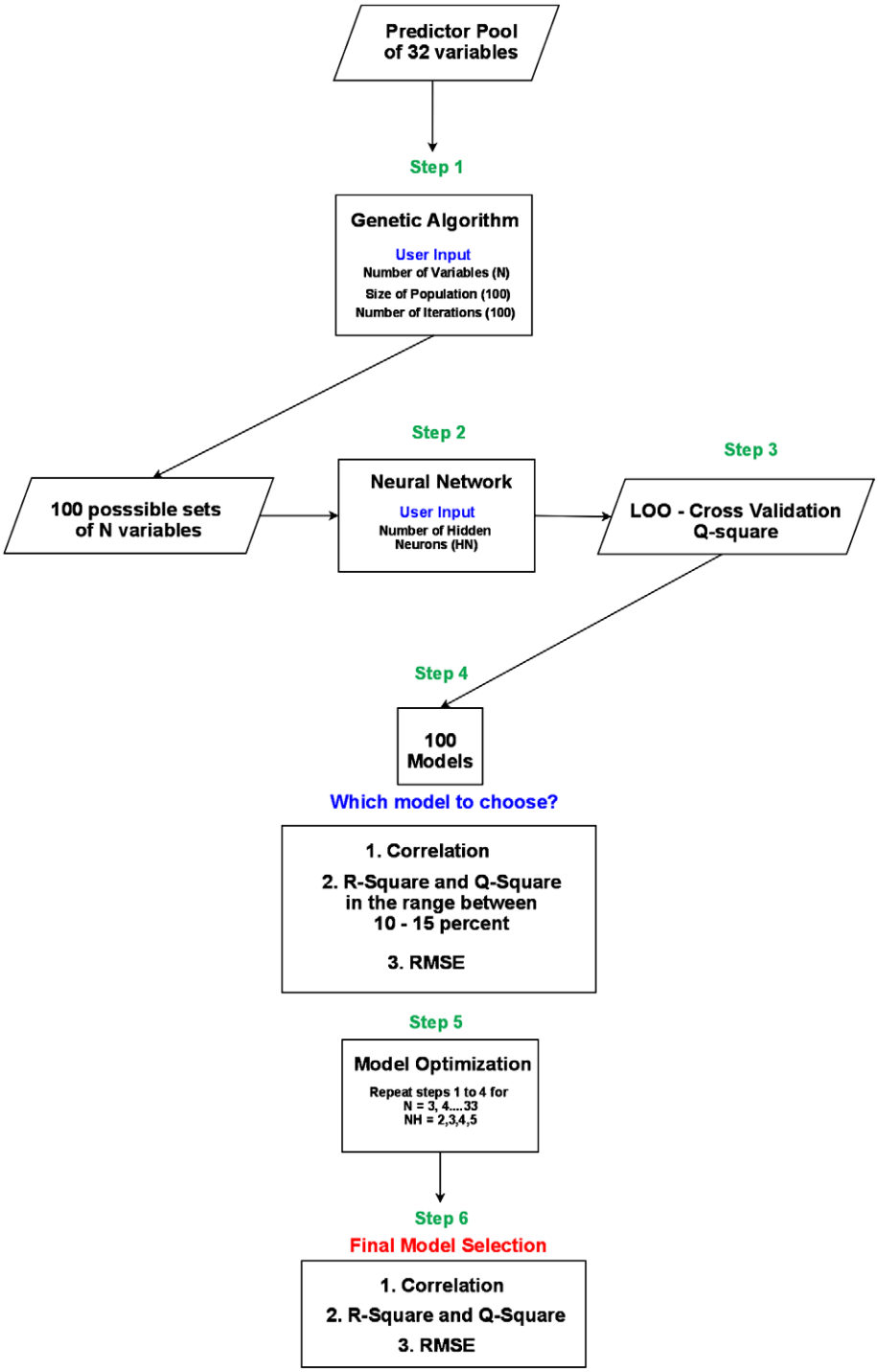

Fig. 4. A flow chart describing the workflow of modeling and optimization techniques used to select the final model.

the final model followed by cross-validation results in the Section 3.3. Section 3.4 also provides a comparison of the model performance of a CNN and an OLS model with the same selected predictors.

3.1. Model selection

The main functions and packages used to build the neural network model were *nnet* function from the library "nnet" and modified genetic algorithm *rbga* function from the library "genalg" of *R* statistical software program. The steps involved are as follows:

3.1.1. Choosing a subset of predictor variables

The predictor pool consisting of 32 variables were passed to the modified *rbga* function. The number of variables (*N*), population size (100), and iterations (1000) were specified. The output were 100 possible subsets of predictor variables of size *N*, say, $Vars_{1 \text{ to } 100}$.

3.1.2. Ensemble of neural network Models

Using the *nnet* function, 100 CNN models were built with $Vars_{1 \text{ to } 100}$ with a user specified number of hidden neurons (*HN*). For each of the 100 CNN model, the response variable was the number

of infected dead birds and the predictor variables were one of the sets of $Vars_{1\ to\ 100}$. In addition, the R^2 and the RMSE values were calculated for each of the 100 models and stored for final model selection.

3.1.3. Cross-validation

The Q^2 values for the 100 CNN models were calculated by the leave one out (LOO) cross-validation method. For each CNN model, the Q^2 value was compared with the associated R^2 value.

3.1.4. Selection of best N-predictor model

Among the 100 CNN models of N predictor variables, the best N-predictor model was chosen based on two criteria: (1) low value of RMSE and (2) R^2 and Q^2 values within the range of 10–15%.

3.1.5. Model optimization

The steps from 1 to 4 are repeated for variable sizes $N = 3, \ldots, 32$ and numbers of hidden neurons f$HN = 2, 3, 4, 5$. Here, we have restricted the number of hidden neurons to five to minimize over fitting. After completing all the optimization runs, we obtained the best 3-predictor model, 4-predictor model, 5-predictor model and so on.

3.1.6. The final model selection

The final model was then chosen from these best N-predictor models by comparing their RMSE and the difference between the R^2 and Q^2 values. The final model had the lowest difference between the R^2 and Q^2 values and low RMSE value. Fig. 4 describes the steps in a flow diagram.

Fig. 5 shows the change in RMSE values calculated from the CNN models with an additional increase of predictor variables (up to 14 variables) and an additional increase of hidden neurons (up to five hidden neurons). The number of variables is shown on the x-axis and RMSE values on the y-axis. The lines denote models with different numbers of hidden neurons. We reported the RMSE values up to 14 variables because beyond $N = 14$ the RMSE's were noisy, with no distinct trend. The behavior of the RMSE values was similar with additional increases of hidden neurons beyond five. The first sharp decline in the RMSE value was identified at the model with five input variables. Beyond this, with an additional input of a risk factor, the RMSE values started to increase and showed inconsistencies, with irregular *peaks* and *dips*. This overall trend of RMSE values was similar for all other models with hidden neurons from 2 to 5.

Fig. 6 shows the difference between the R^2 and Q^2 values. Here the number of variables is shown on the x-axis, the difference between R^2 and Q^2 on the y-axis, and the lines denote models with increasing number of hidden neurons. The figure shows that the CNN model with two hidden layer neurons and five predictor variables had the lowest difference between them. Thus, based on the RMSE values (Fig. 5) and the difference between R^2 and Q^2 values (Fig. 6), 5-2-1 CNN architecture model with five input neurons (predictor variables), two hidden neurons, and one output neuron was selected as the final neural network WNV analysis model.

3.2. Description of the selected model

The 5-2-1 CNN model had a low RMSE value of 1.78 and the lowest difference (13%) between R^2 (0.75) and Q^2 (0.62) values. The model had five input neurons, each corresponding to one of the predictor variables chosen by the genetic algorithm, two hidden neurons, and one output neuron, which stored the predicted number of WNV infected dead birds for a zip code. The predictor variables selected for this model were distance to bogs (miles), distance to lakes (miles), daily maximum temperature (F), age of houses (years), and percentage of developed medium density land cover class (Fig. 7). The histogram of the observed data shows a negatively skewed distribution with higher frequencies of zip codes with no reports, followed by 0–2 cases of dead bird reports, and so on (Fig. 8). As expected in a negatively skewed distribution, the histogram has a thin and elongated tail indicating lower frequencies of higher dead bird counts. The CNN predicted histogram, in overall, shows a similar pattern, with strong correlations between the observed and predicted values for fewer numbers of infected dead birds and relatively weaker correlations for higher counts (Fig. 8). Out of 93 zip codes with observed 0–2 dead bird cases, the CNN model correctly predicted the number of infected dead birds for 74 zip codes. The spatial distribution of the observed and predicted numbers of infected dead birds also shows a similar pattern (Fig. 9).

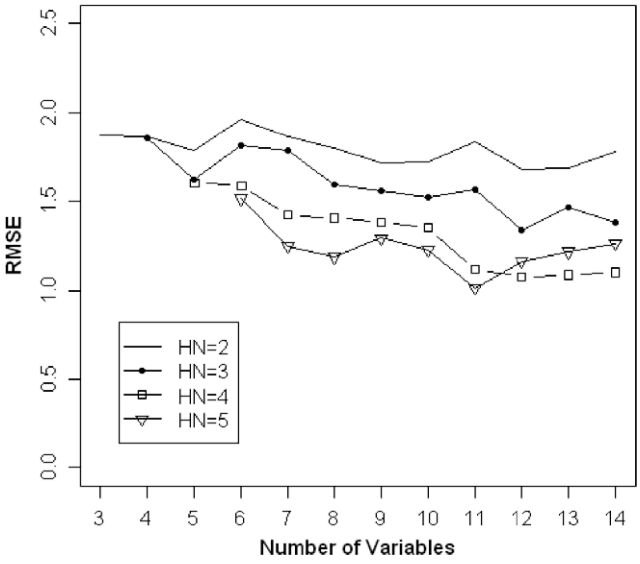

Fig. 5. Behavior of RMSE values with an additional increase of predictor variable and an additional increase of hidden neurons.

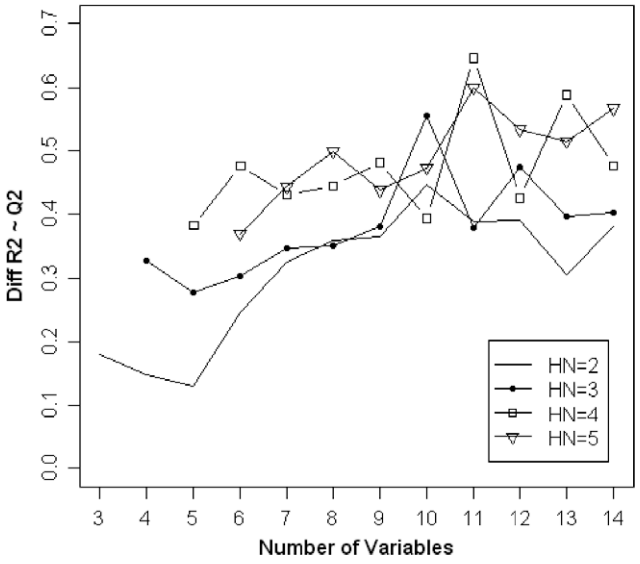

Fig. 6. Difference between R^2 and Q^2 values with an additional increase of predictor variables and an additional increase of hidden neurons.

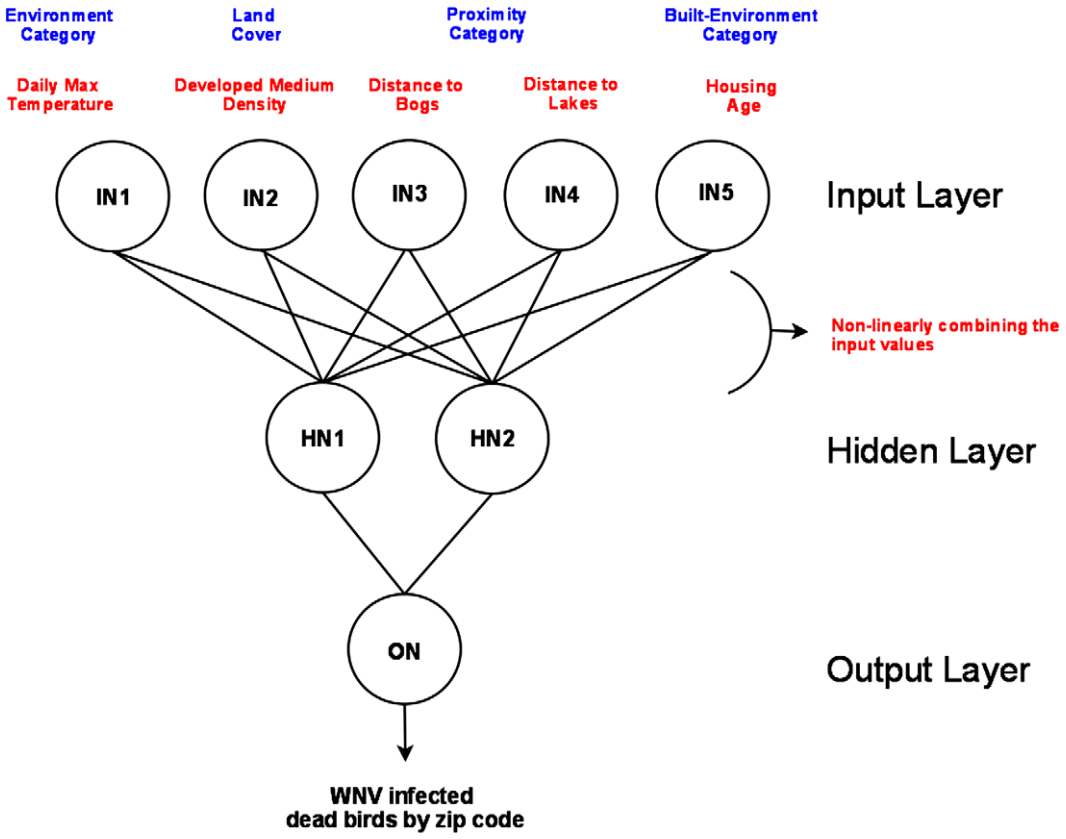

Fig. 7. Structure of West Nile virus analysis model using feed-forward neural network Algorithm.

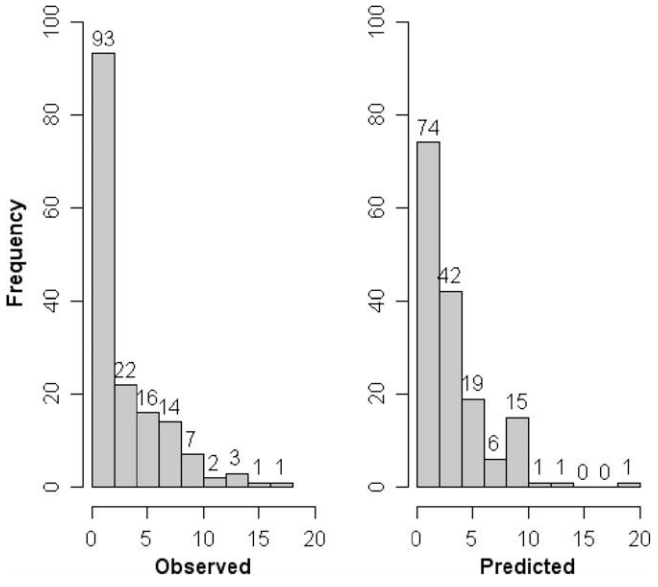

Fig. 8. Histograms of observed and predicted number of West Nile virus infected dead birds by zip codes in 2006.

3.3. Cross-validation

For cross-validation, two external datasets were considered. These new datasets were the observed number of WNV infected dead birds in the year 2003 and the observed number of WNV infected dead birds in the year 2007. The data from the year 2003 was selected for retrospective prediction, with 285 infected dead birds. A total of 60 infected bird cases from the year 2007 were chosen for prospective prediction. The use of these datasets will assess the predictive capability of the CNN model in two different scenarios with relatively higher (285) and lower (60) counts of infected dead birds. The descriptive statistics of the variables in the two datasets are shown in Table 3 and Table 4. Neural network models were built for both the datasets with the same 5-2-1 architecture of the WNV analysis model. The resultant RMSE, R^2, and observed versus predicted plots from the external datasets were analyzed to assess the predictive capability of the WNV model.

The RMSE and R^2 values of the 2003 dataset obtained from the CNN model were 1.01 and 0.65, respectively. Both of the values were acceptable, given that the R^2 value was high and the RMSE value was close to 10% of the range of infected numbers of dead birds (0–10) in 2003. The histograms of observed verses predicted values are shown in Fig. 10. Overall, both the observed and the predicted histograms showed similar patterns of negatively skewed distributions with higher frequency of fewer dead bird counts and vice versa. In the original data, 92 zip codes reported no bird cases, and 79 (86%) of such zip codes were predicted correctly by the CNN model. Further, the CNN model was almost 100% accurate in predicting the dead bird counts for zip codes with number greater than eight. However, the predictions were relatively weaker for the middle values, especially for zip codes with 3–4 dead birds. In the observed distribution, there were similar frequencies of zip codes with 3, 4, and 5 bird counts, but the predicted histogram showed a bimodal distribution with 38 zip codes with 3–4 dead bird cases. In addition, Fig. 11 also shows a similar spatial pattern of observed versus predicted cases with generally higher numbers of infected dead birds reported in the zip codes around the Twin Cities of Minneapolis and Saint Paul.

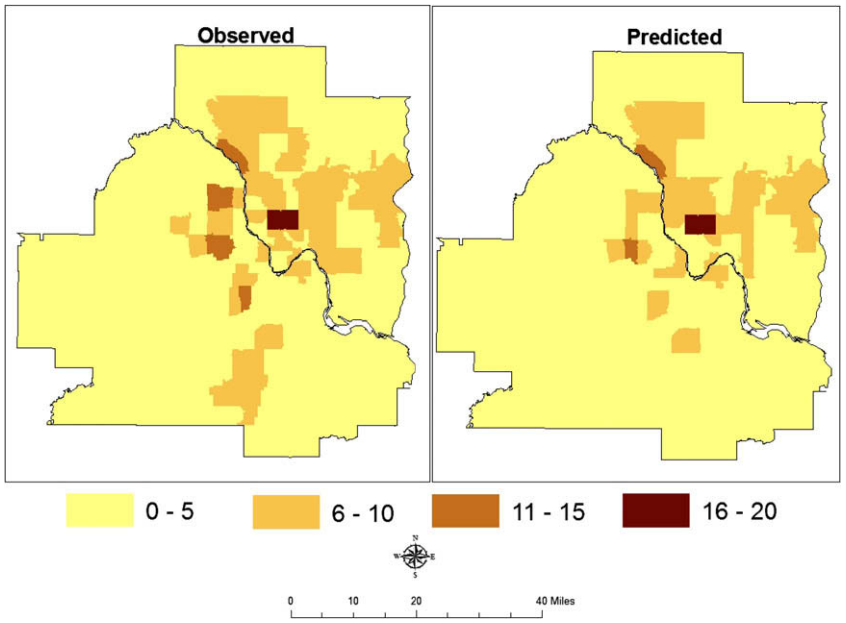

Fig. 9. Spatial distribution of observed and predicted number of West Nile virus infected dead birds by zip codes in 2006.

Table 3
Descriptive statistics of the 2003 dataset.

Variables	Min.	1st quartile	Median	Mean	3rd quartile	Max.
Observed Y						
Bird_2003	0	0	1	2	3	10
Predictors						
Age of houses (years)	0	22	35	39	48	88
Distance to bogs (miles)	661.40	2834.70	5654.90	6995.50	8849.30	25297.80
Distance to lakes (miles)	60.59	301.48	440.46	524.30	646.11	1871.95
Dev, medium density (%)	0.62	4.41	17.03	18.80	30.06	64.28
Max daily temperature (*F*)	69.16	72.59	80.57	76.53	84.08	91.29

Table 4
Descriptive statistics of the 2007 dataset.

Variables	Min.	1st quartile	Median	Mean	3rd quartile	Max.
Observed Y						
Bird_2007	0	0	0	0	1	6
Predictors						
Age of houses (years)	1	26	38	42	51	91
Distance to bogs (miles)	661.40	2834.70	5654.90	6995.50	8849.30	25297.80
Distance to lakes (miles)	60.59	301.48	440.46	524.30	646.11	1871.95
Dev, medium density (%)	0.62	4.41	17.03	18.80	30.06	64.28
Max daily temperature (*F*)	67.68	71.69	73.87	78.31	86.55	91.86

The RMSE and R^2 values of the 2007 dataset were 0.71 and 0.74, respectively. Both, a spatial distribution (Fig. 12) and spatial distribution (Fig. 14) of the observed and the predicted number of infected dead birds show similar patterns. In Fig. 14, both histograms have similar distributions, with the highest frequency of zip codes with no cases, followed by zip codes with one case, and so on. Out of the 114 zip codes with no cases, 110 (96%) were predicted correctly. The strong correlation between the observed and predicted values observed in the histograms is also reflected in the similar spatial distribution shown by the choropleth maps in Fig. 13. Thus, the overall similar trends of a spatial (histograms) and spatial (choropleth maps) distributions of the observed and the predicted values, the higher R^2, and the lower RMSE values for both external datasets indicated that the CNN model had good predictive capabilities.

3.4. Comparison of CNN and OLS models

To compare with the results of the selected WNV neural network model, an OLS model was build with same specifications. The OLS model had an *F*-statistic of 32.2 (for 5 and 153 degrees of freedom), which was much greater than the critical value of 1.96 ($\alpha = 0.05$). The model was thus statistically valid. The comparison of OLS and CNN models based on RMSE, R^2, and Q^2 are shown in Table 5. Even though the difference between the $R^2$and Q^2values was the lowest in the case of the OLS model, the CNN model

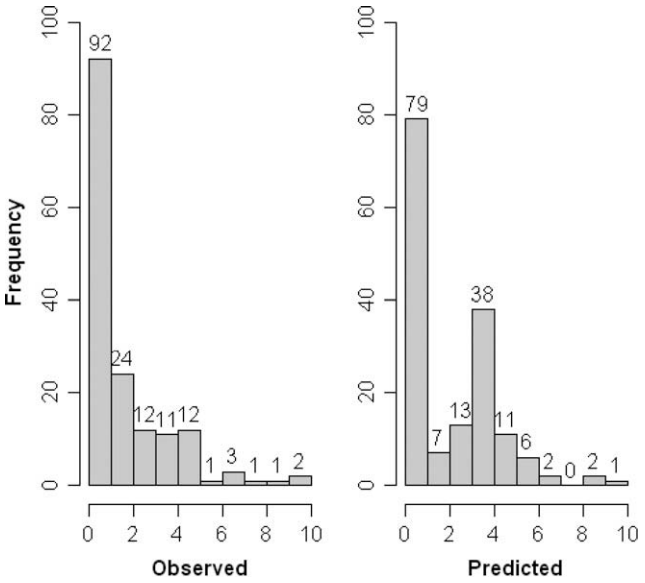

Fig. 10. Histograms of observed and predicted number of West Nile virus infected dead birds by zip codes in 2003.

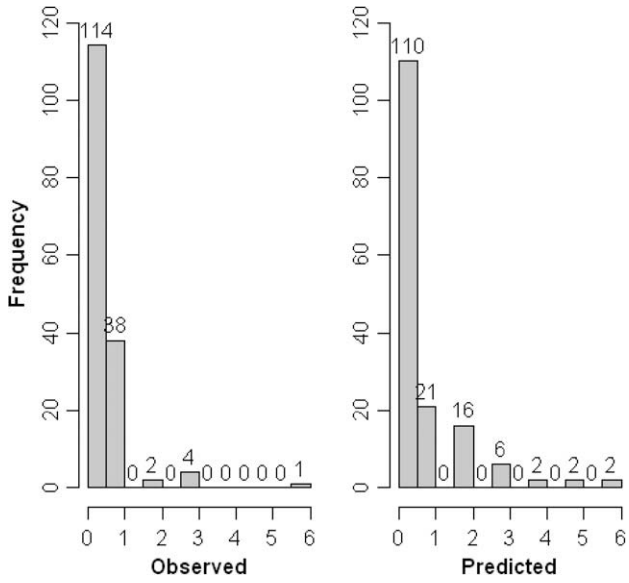

Fig. 12. Histograms of observed and predicted number of West Nile virus infected dead birds by zip codes in 2007.

performed better than the OLS model in terms of other indicators, such as lower RMSE and higher values. Moreover, the internal cross-validation result (13%) was within the acceptable range of 10–15%. In Fig. 14, the histogram of the observed data showed a negatively skewed distribution. The CNN predicted histogram in general showed similar patterns, with strong correlation between the observed and predicted values for fewer numbers of infected dead birds and relatively weaker correlations for higher counts. On the other hand, the distribution of the predicted values from the OLS model was significantly different from the observed data. The OLS predicted distribution was bimodal, with the highest frequency for the 0 case count followed by the 5–6 case counts. Out of 93 zip codes with 0–2 dead bird cases, only 58 (62%) were predicted correctly by the OLS model. On the other hand, the CNN model correctly predicted 74 of such zip codes. Another striking difference was that the OLS model underestimated the response variable. In the original dataset, the highest number of infected dead birds for a particular zip code was 17 cases. However, the highest predicted value from the OLS model was only 9 cases, and that of the CNN model was as high as 19 cases.

Thus, the comparative analysis of the OLS and CNN models based on the RMSE, R^2, Q^2, (Table 5) and histograms (Fig. 14) of the observed and predicted values indicated that the 5-2-1 CNN model was superior to the OLS model in terms of quality and predictability. It also suggested that computational neural network models are better suited to capture nonlinear relationships between the predictors and response variable, such as involved in the dynamics of WNV infection.

4. Interpretation of West Nile virus analysis model

This section reports the results obtained from the broad interpretation of the WNV neural network model. The broad interpreta-

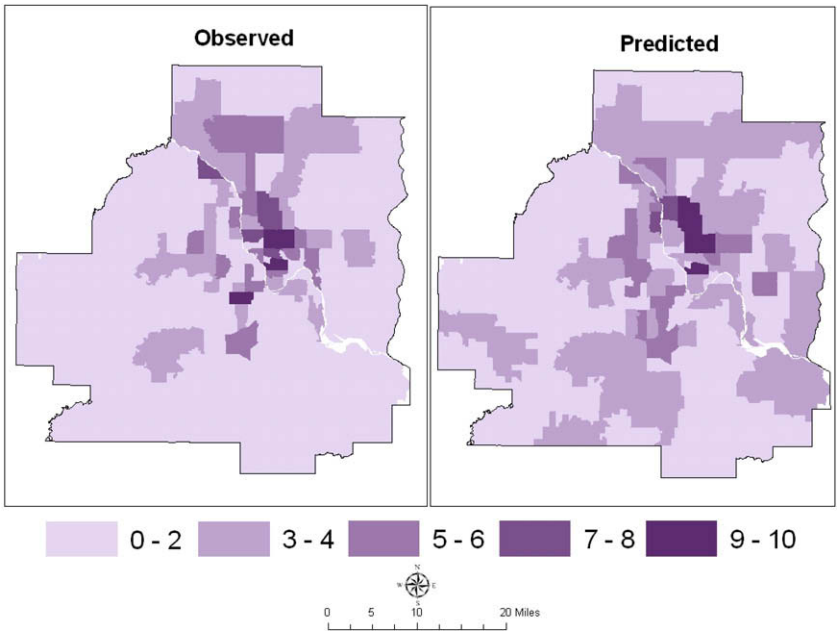

Fig. 11. Spatial distribution of observed and predicted number of West Nile virus infected dead birds by zip codes in 2003.

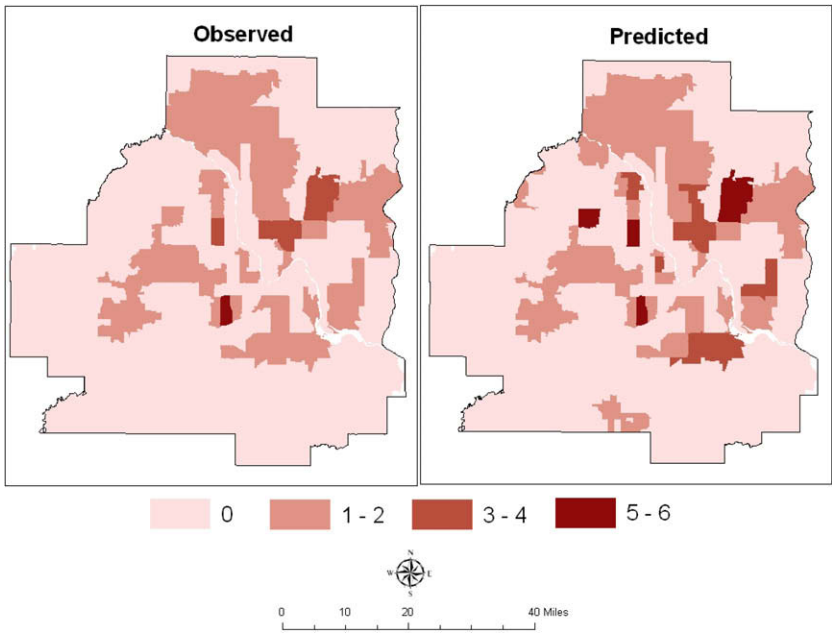

Fig. 13. Spatial distribution of observed and predicted number of West Nile virus infected dead birds by zip codes in 2007.

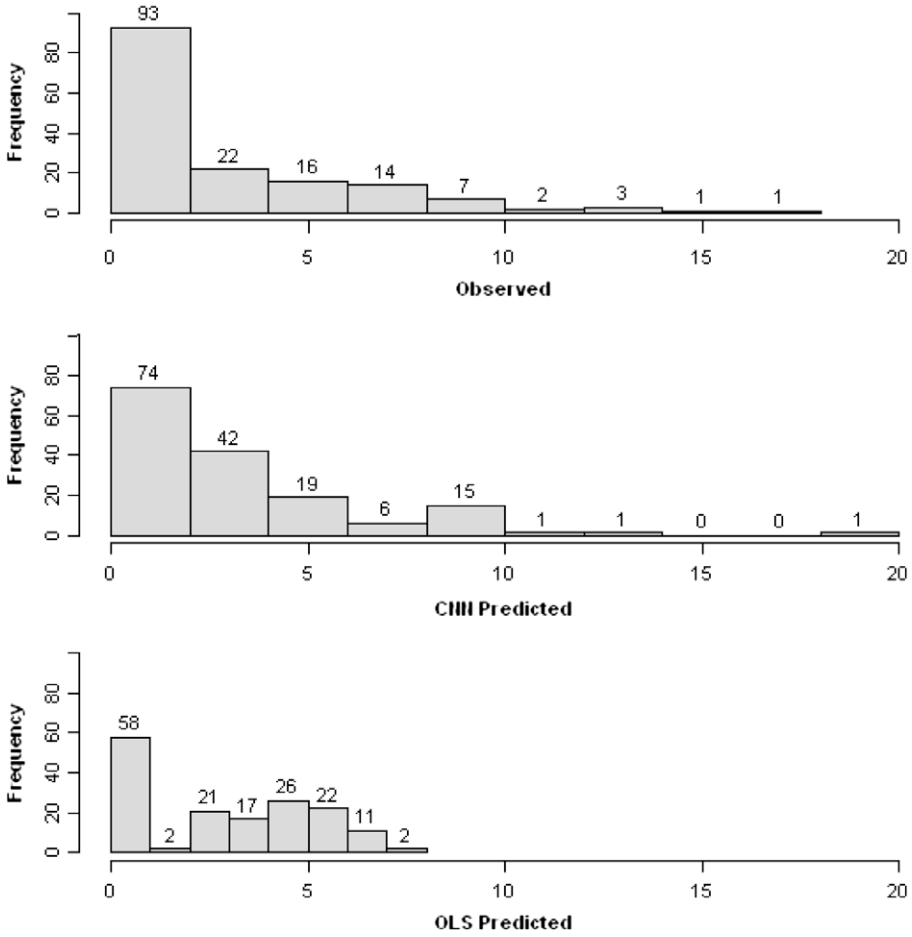

Fig. 14. Histograms showing the comparison of observed, CNN predicted, and OLS predicted values of West Nile virus infected dead birds by zip codes in 2006.

tion approach measures the importance of predictor variables included in the model in terms of their contribution to the response variable (see Section 2.4 for details).

The increase in RMSE values for the predictors in the CNN model is reported in Table 6. The fourth column in the table represents the increase in RMSE due to the scrambling of the corresponding

Table 5
Comparison of OLS and CNN models with the best subset of predictors.

Indicators	OLS	CNN
RMSE	2.56	1.78
R^2	0.51	0.75
Q^2	0.4	0.62
Difference	0.11	0.13

Table 6
Increase in RMSE due to scrambling of individual predictor variables. The CNN architecture is 5-2-1 with a base RMSE of 1.78.

Scrambled predictors	Description	RMSE	Difference
Tmax	Daily max temperature	3.308	1.528
Bog	Distance to bogs	2.673	0.893
HoAge	Housing age	2.446	0.666
Med. density	Dev, medium density	2.315	0.535
Lake	Distance to lake	2.012	0.232

predictor variable over the base RMSE of 1.78. It was evident that scrambling some descriptors led to larger increases, whereas others led to negligible increases in the RMSE. The information contained in Table 6 is more easily seen in the predictor importance plots shown in the Fig. 15. This figure plots the increase in RMSE for each input factor in decreasing order.

Considering Table 6 and Fig. 15, we can say that the most important risk factor contributing to the occurrence of WNV infection in birds in the TCMA was the maximum daily temperature. This risk factor represents an important environmental determinant associated to the occurrence of WNV incidences and is heavily documented in the literature (Adlouni, Beaulieu, Ouarda, Gosselin, & Saint-Hilaire, 2007; Bolling, Moore, Anderson, Blair, & Beaty, 2007; Bouden, Moulin, & Gosselin, 2008; DeGroote, Sugumaran, Brend, Tucker, & Bartholomay, 2008; Gibbs et al., 2006; Gleiser et al., 2007; Hayes et al., 2005; Landesman, Allan, Langerhans, Knight, & Chase, 2007; O'Leary et al., 2004; Ozdenerol, Bialkowska-Jelinska, & Taff, 2008; Ruiz et al., 2004; Savage et al., 2006; Theophilides, Ahearn, Binkowski, Paul, & Gibbs, 2006; Theophilides, Ahearn, Grady, & Merlino, 2003; Vaidyanathan & Scott, 2006; Zou, Miller, & Schmidtman, 2007). The second most important risk factor was distance to bogs. This study made a significant contribution to the literature of WNV by investigating the relationship between WNV disease occurrences and *proximity* to hypothesized risk factors, such as bogs and lakes. Other studies have quantified these risk factors as percentage of lake area or presence and absence of a lake in an analysis unit (Cooke, Grala, & Wallis, 2006; Ezenwa et al., 2007; Leblond et al., 2007; Reisen, Lundstorm, Scott, & Eldridge, 2000; Schafer, Lundstorm, Pferrer, & Lundkvist, 2004; Williot, 2004). In contrast, we attempted to analyze an important question, i.e., how proximity to a particular risk factor influences the risk of WNV. Bogs are among the major wetland types classified by MMCD as potential breeding sites for mosquitoes. A bog is described as a wetland with vegetation including moss, sedges, cotton grass, and shrubs, poorly drained soil, waterlogged, and supporting a covering thick mat of plant residues. In Fig. 15, the significant gap between these two risk factors along the *x*-axis (relative increase in RMSE) qualitatively indicated that maximum daily temperature was contributing much more to the predicted output than the distance to bogs.

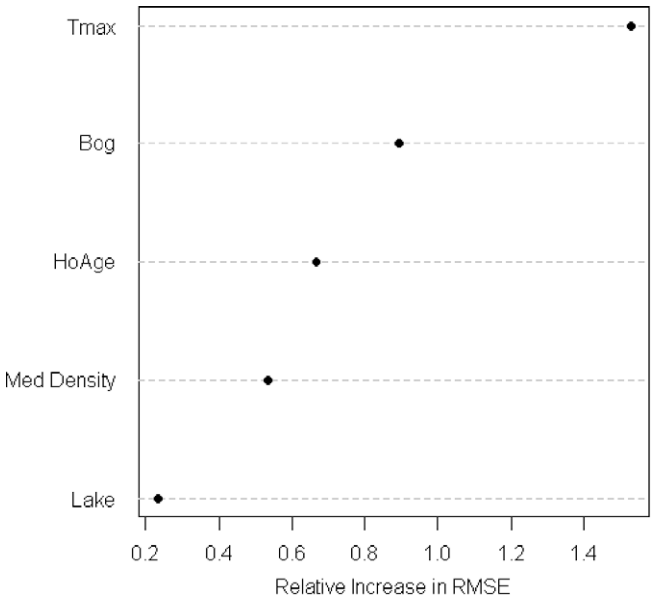

Fig. 15. Importance plot for the 5-2-1 West Nile virus CNN model.

The next two important factors were housing age and percentage of developed medium density land cover, both of which belonged to the built-environment category of hypothesized risk factors. The land cover class is defined as areas with mixture of constructed materials and vegetation, impervious surface accounting for 20–49% of total cover, and is mostly occupied by single-family housing with typical grassy backyards and swimming pools. The selection of these two risk factors and their level of importance were in tune with the urban-centric nature of WNV transmission in the TCMA. The insignificant separation between these two variables along the *x*-axis also indicated that these two factors were probably playing similar roles in the predictive ability of the CNN model. The last risk factor in the order of importance was distance to lakes.

5. Discussion and conclusion

This paper presented an intrinsically nonlinear approach (neural network) to predict the number of WNV infected dead birds for the entire seven county metropolitan area of Minnesota. The specific goals of this work were first, to describe in detail the procedure involved in building a CNN model incorporating the nonlinearity between disease occurrence and associated risk factors, and second, to propose an interpretation method that would rank the selected risk factors in order of their importance.

For the first objective of this paper, GA was used to search the predictor space for the best or optimal subset of predictor variables, which were then included in a CNN model. The final model selection was based on a combination of lower RMSE values, higher R^2, and lower difference between R^2 and Q^2 values. This procedure resulted in a CNN model with 5-2-1 architecture, as the suitable model for WNV analysis. The predictor variables included were distance to bogs, distance to lakes, daily maximum temperature, housing age, and percentage of developed medium density land cover class. Even though the OLS model with the same specifications was statistically significant (*F*-statistic), the RMSE was significantly higher and the R^2 lower than the values obtained from the CNN model. The observed versus predicted histograms and the choropleth maps also suggested that the CNN model had good predictive power when modeling a complex health outcome, such as WNV. In addition, the cross-validation results with new datasets confirmed the superior predictive performance of the model.

Another focus of this paper was to measure the relative roles of the selected risk factors on the predictive capability of the WNV neural network model using the proposed broad interpretation

technique. The interpretation method, which is essentially a sensitivity analysis of CNN models, was able to rank the relative importance of the selected five risk factors. The risk factors in descending order of importance were average daily maximum temperature, distance to bogs, housing age, medium density land cover class, and distance to lakes. The risk factor importance plot also allowed the easy visualization of the ranked predictors. Apart from the quantitative rankings, the importance plot also provided a qualitative view of how important a given predictor was relative to others by looking at the difference between RMSE, represented on the *x*-axis. If there was a large gap or separation, for example, between average daily maximum temperature and distance to bogs, one could determine that temperature plays a much more significant role than distance to bogs in the model's predictive power.

Even though the broad interpretation approach provided an easy measure to rank the selected risk factors, its abilities were subject to some key limitations. It only provided relative predictor *importance* and did not allow the user to elucidate exactly how a given risk factor *affected* the model output. That is, the methodology did not indicate the relationships (direction of correlation) between the input risk factors and occurrence of WNV incidences. Further, the interpretation method did not have the capacity to explain in detail the epidemiological understanding of the interrelationships among the risk factors and WNV disease incidences identified by the CNN model. This *black box* nature or lack of interpretability of neural network approaches often forces researchers to use CNN models as predictive tools rather than as explanatory tools. This is in contrast to linear models, which can be interpreted in a simple manner but have poor predictive ability. A follow up to this paper (currently under review) addresses this important research need. The future work proposes a *'detailed'* interpretation technique, which will extract the relationships between WNV disease occurrence and the selected risk factors encoded in the weights and biases of a CNN model.

References

Adlouni, S. E., Beaulieu, C., Ouarda, T. B., Gosselin, P. L., & Saint-Hilaire, A. (2007). Effects of climate on West Nile virus transmission risk used for public health decision-making in Quebec. *International Journal of Geographic Information Science, 6*(40).

Altman, D. G., & Anderson, P. K. (1989). Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine, 8*, 771–783.

Bolling, B. G., Moore, C. G., Anderson, S. L., Blair, C. D., & Beaty, B. J. (2007). Entomological studies along the Colorado front range during a period of intense West Nile virus activity. *Journal of the American Mosquito Control Association, 23*(1), 37–46 (Article).

Bouden, M., Moulin, B., & Gosselin, P. L. (2008). The geosimulation of West Nile virus propagation: A multi-agent and climate sensitive tool for risk management in public health. *International Journal of Health Geographics, 7*(35).

Bowman, C., Gumel, A. B., Driessche, P. v. d., Wu, J., & Zhu, H. (2005). A mathematical model for assessing control strategies against West Nile virus. *Bulletin of Mathematical Biology, 67*, 1107–1133.

Brown, H. E., Childs, J. E., Diuk-Wasser, M. A., & Fish, D. (2008). Ecological factors associated with West Nile virus transmission, northeastern United States. *Emerging Infectious Diseases, 14*(10), 1539–1545 (Article).

Brownstein, J. S., Rosen, H., Purdy, D., Miller, J. R., Merlino, M., & Mostashari, F. (2002). Spatial analysis of West Nile virus: Rapid risk assessment of an introduced vector-borne zoonosis. *Vector Borne Zoonotic Disease, 2*, 157–164.

CDC. (1999). Outbreak of West Nile like viral encephalitis. New York: Centers for Disease Control and Prevention.

Cooke, W. I., Grala, K., & Wallis, R. (2006). Avian GIS models signal human risk for West Nile virus in Mississippi. *International Journal of Geographic Information Science, 5*(36).

Cooke, W. H., Katarzyna, G., & Wallis, R. C. (2006). Avian GIS models signal human risk for West Nile virus in Mississippi. *International Journal of Health Geographies, 5*(36).

David, S., Mak, S., MacDougall, L., & Fyfe, M. (2007). A bird's eye view: Using geographic analysis to evaluate the representativeness of corvid indicators for West Nile virus surveillance. *International Journal of Health Geographics, 6*(1), 3.

DeGroote, J. P., Sugumaran, R., Brend, S. M., Tucker, B. J., & Bartholomay, L. C. (2008). Landscape, demographic, entomological, and climatic associations with human disease incidence of West Nile virus in the state of Iowa, USA. *International Journal of Health Geographics, 7*(19).

Diuk-Wasser, M. A., Brown, H. E., Andreadis, T. G., & Fish, D. (2006). Modeling the spatial distribution of mosquito vectors for West Nile virus in Connecticut, USA. *Vector-Borne and Zoonotic Diseases, 6*(3), 283–295 (Article).

Ezenwa, V., Milheim, L., Coffey, M., Godsey, M., King, R., & Guptill, S. (2007). Land cover variation and West Nile virus prevalence: Patterns, processes, and implications for disease control. *Vector Borne Zoonotic Diseases, 7*(2), 173–180.

Fischer, M. M. (1998). Computational neural networks: A new paradigm for spatial analysis. *Environment and Planning A, 30*(10), 1873–1891.

Forrest, S. (1993). Genetic algorithms: Principles of natural selection applied to computation. *Science, 261*, 872–878.

Gibbs, S. E. J., Wimberly, M. C., Madden, M., Masour, J., Yabsley, M. J., & Stallknecht, D. E. (2006). Factors affecting the geographic distribution of West Nile virus in Georgia, USA: 2002–2004. *Vector-Borne and Zoonotic Diseases, 6*(1), 73–82.

Gleiser, R. M., Mackay, A. J., Roy, A., Yates, M. M., Vaeth, R. H., Faget, G. M., et al. (2007). West Nile virus surveillance in East Baton Rouge Parish, Louisiana. *Journal of the American Mosquito Control Association, 23*(1), 29–36 (Article).

Golbraikh, A., & Tropsha, A. (2002). Beware of Q2. *Journal of Molecular Graphics and Modeling, 20*, 269–276.

Guha, R. (2005). *Methods to improve the reliability, validity, and interpretability of QSAR Models*. Pennsylvania State University, State College, PA.

Guha, R., & Jurs, P. (2005). Interpreting computational neural network QSAR Models: A measure of descriptor importance. *Journal of Chemical Information and Modeling, 45*(3), 800–806.

Guha, R., Stanton, D., & Jurs, P. (2005). Interpreting computational neural networks QSAR Models: A detailed interpretation of the weights and biases. *Journal of Chemical Information and Modeling, 45*(4), 1109–1121.

Hayes, E. B., Komar, N., Nasci, R. S., Montgomery, S. P., O'Leary, D. R., & Campbell, G. L. (2005). Epidemiology and transmission dynamics of West Nile virus disease. *Emerging Infectious Disease, 11*(8), 1173–1176.

Haykin, S. (2001). *Neural networks*. Singapore: Pearson Education.

Johnson, G. D. (2008). Prospective spatial prediction of infectious disease: Experience of New York state (USA) with West Nile virus and proposed directions for improved surveillance. *Environmental and Ecological Statistics, 15*(3), 293–311 (Proceedings Paper).

LaBeaud, A. D., Gorman, A. M., Koonce, J., Kippes, C., McLeod, J., Lynch, J., et al. (2008). Rapid GIS-based profiling of West Nile virus transmission: Defining environmental factors associated with an urban–suburban outbreak in Northeast Ohio, USA. *Geospatial Health, 2*(2), 215–225 (Article).

Landesman, W., Allan, B., Langerhans, R., Knight, T., & Chase, J. (2007). Inter-annual associations between precipitation and human incidence of West Nile virus in the United States. *Vector Borne Zoonotic Diseases, 7*(3), 337–343.

Leblond, A., Sandoz, A., Lefebvre, G., Zeller, H., & Bicout, D. J. (2007). Remote sensing based identification of environmental risk factors associated with West Nile disease in horses in Camargue, France. *Preventive Veterinary Medicine, 79*(1), 20–31 (Article).

Lian, M., Warner, R., Alexander, J., & Dixon, K. (2007). Using geographic information systems and spatial and space-time scan statistics for a population-based risk analysis of the 2002 equine West Nile epidemic in six contiguous regions of Texas. *International Journal of Health Geographics, 6*(1), 42.

Mennis, J., & Diansheng, G. (2009). Spatial data mining and geographic knowledge discovery—An introduction computers. *Environment and Urban Systems, 33*(6), 403–408.

Mongoh, M. N., Khaitsa, M. L., & Dyer, N. W. (2007). Environmental and ecological determinants of West Nile virus occurrence in horses in North Dakota, 2002. *Epidemiology and Infection, 135*(1), 57–66 (Article).

Mostashari, F., Martin, K., Hartman, J. J., Miller, J. R., & Kulasekera, V. (2003). Dead bird clusters as an early system for West Nile virus activity. *Emerging Infectious Diseases, 9*(6), 641–646.

Murray, K., Baraniuk, S., Resnick, M., Arafat, R., Kilborn, C., Cain, K., et al. (2006). Risk factors for encephalitis and death from West Nile virus infection. *Epidemiology and Infection, 134*(6), 1325–1332 (Article).

O'Leary, D., Marfin, A., Montgomery, S., Kipp, A. M., Lehman, J., & Biggerstaff, B. (2004). The epidemic of West Nile virus in the United States, 2002. *Vector Borne Zoonotic Disease, 4*, 61–70.

Ozdenerol, E., Bialkowska-Jelinska, E., & Taff, G. N. (2008). Locating suitable habitats for West Nile virus-infected mosquitoes through association of environmental characteristics with infected mosquito locations: A case study in Shelby County, Tennessee. *International Journal of Health Geographics, 7*(12).

Pradier, S., Leblond, A., & Durand, B. (2008). Land cover, landscape structure, and West Nile virus circulation in southern France. *Vector-Borne and Zoonotic Diseases, 8*(2), 253–263 (Article).

Reisen, W. K., Lundstorm, J. O., Scott, T. W., & Eldridge, B. F. (2000). Patterns of avian seroprevalence to Western Equine Encephalomyelitis and Saint Louis Encephalitis viruses in California, USA. *Journal of Medical Entomology, 37*(507–527).

Roberts, R. S., & Foppa, I. M. (2006). Prediction of equine risk of West Nile Virus infection based on dead bird surveillance. *Vector-Borne and Zoonotic Diseases, 6*(1), 1–6 (Article).

Ruiz, M. O., Tedesco, C., McTighe, T. J., Austin, C., & Kitron, U. (2004). Environmental and social determinants of human risk during a West Nile virus outbreak in the greater Chicago area, 2002. *International Journal of Health Geographics, 3*(8), 2–11.

Ruiz, M. O., Walker, E. D., Foster, E. S., Haramis, L. D., & Kitron, U. D. (2007). Association of west Nile virus illness and urban landscapes in Chicago and Detroit. *International Journal of Health Geographics, 6*(10), 1–11.

Savage, H. M., Anderson, M., Gordon, E., McMillen, L., Colton, L., Charnetzky, D., et al. (2006). Oviposition activity patterns and West Nile virus infection rates for members of the *Culex pipiens* complex at different habitat types within the hybrid zone, Shelby County, TN, 2002 (Diptera: Culicidae). *Journal of Medical Entomology, 43*(6), 1227–1238 (Article).

Schafer, M. L., Lundstorm, J. O., Pferrer, M., & Lundkvist, E. (2004). Biological diversity verses risk for mosquito nuisance and disease transmission in constructed wetlands in southern Sweden. *Medical Veterinary and Entomology, 18*, 256–267.

Shad, R., Mesgari, M. S., Abkar, A., & Shad, A. (2009). Predicting air pollution using fuzzy genetic linear membership kriging in GIS Computers. *Environment and Urban Systems, 33*(6), 472–481.

Shaman, J., Day, J. F., & Stieglitz, M. (2005). Drought-induced amplification and epidemic transmission of West Nile virus in Southern Florida. *Journal of Medical Entomology, 42*(2), 134–141.

Theophilides, C. N., Ahearn, S. C., Binkowski, E. S., Paul, W. S., & Gibbs, K. (2006). First evidence of West Nile virus amplification and relationship to human infections. *International Journal of Geographical Information Science, 20*(1), 103–115.

Theophilides, C. N., Ahearn, S. C., Grady, S., & Merlino, M. (2003). Identifying West Nile virus risk areas: The dynamic continuous-area space-time system. *American Journal of Epidemiology, 157*(9), 843–854 (Article).

Vaidyanathan, R., & Scott, T. W. (2006). Seasonal variation in susceptibility to West Nile virus infection in *Culex pipiens* (L.) (Diptera : Culicidae) from San Joaquin County, California. *Journal of Vector Ecology, 31*(2), 423–425 (Article).

Williot, E. (2004). Restoring nature, without mosquitoes? *Restoring Ecology, 12*, 147–153.

Yiannakoulias, N. W., Schopflocher, D. P., & Svenson, L. W. (2006). Modelling geographic variations in West Nile virus. *Canadian Journal of Public Health-Revue Canadienne De Sante Publique, 97*(5), 374–378 (Article).

Yiannakoulias, N. W., & Svenson, L. W. (2007). West Nile virus: Strategies for predicting municipal-level infection. *Annals of New York Academy of Sciences, 1102*, 135–148.

Zou, L., Miller, S., & Schmidtman, E. (2007). A GIS tool to estimate West Nile virus risk based on a degree-day model. *Environmental Monitoring and Assessment, 126*, 413–420.