DYNAMIC NETWORK MODELS OF A CONTINENTAL EPIDEMIC: SOYBEAN RUST IN THE USA

by

SWETA SUTRAVE

B.E., Osmania University, India, 2007

A THESIS

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Electrical and Computer Engineering
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas
2010

Approved by:

Co-Major Professor
Caterina Scoglio

Approved by:

Co-Major Professor
Karen Garrett

Copyright

Abstract

With rapid global movement of epidemics, research efforts to characterize dynamics of epidemics have gained much focus. Traditional epidemiological models have focused on only temporal components of epidemics. Development of spatio-temporal models proved to be a notable achievement in epidemiology. Network-based epidemiological models enable better handling of spatial and temporal components of an epidemic. Early network models considered a binary level of contact between infected entities, which is an idealistic approach. A realistic approach would use weighted edges which signify the level of interaction between the nodes where the edge-weights change over time as a function of environmental factors. Estimation of edge weights from observed time series is a relatively less explored area for network modeling. Dynamic networks make the problem more complicated as edge weights change over time. Estimation of parameters for models describing the edge weights as a function of variables that change in time has the potential to provide better general models. Soybean rust (caused by *Phakopsora pachyrhizi*) is an important disease globally and its occurrence in the US has been studied extensively since its introduction in 2004. Rust is a fungal disease which propagates as a result of the fungal spores being carried by the wind. In this thesis, a network network based model is proposed to predict the intensity of spread of the disease in space and time. This model uses the host abundance and wind data and the observed rust incidence time series to compute the edge-weights. Also, the edge-weights in the model change over time thus following a dynamic approach. In order to cut costs involved with the establishment and maintenance of infection monitoring sites, the effect of removal of monitoring nodes using various strategies has also been analyzed in this thesis. The model has been tested with observed soybean rust data from sentinel plot network from across the United States.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisors, Dr.Karen Garrett and Dr.Caterina Scoglio for their guidance and continued support in my project. I am grateful to Dr.Scott Isard for providing me the Soybean rust data and valuable advice. I would also like to thank the NSF and USDA-APHIS for their support. Finally, I thank Phillip Schumm and Margaret Margosian for their advice.

Chapter 1

Introduction and prior work

As global movement of diseases becomes more rapid, it is increasingly important to develop flexible predictive models to support epidemiology. Significant efforts in modeling of soybean rust began about two decades ago[1]. In order to monitor the disease in the US, a network of sentinel plots was organized by soybean researchers and organizations[2–4]. Network models offer techniques for evaluating epidemic spread across landscapes and habitats[5]. Network model applications in plant disease epidemiology enable better understanding of spatial and temporal components of epidemics in plants[6–9]. In Wang et al. (2003)[10], the authors present a model for spread of virus in a network. They determined a universal epidemic threshold for SI model or a binary contact network with susceptible and infected nodes. They also demonstrate that the topology of the network greatly influences the rate of spread. The authors Pastor-Satorras and Vespignani[11,12] showed the absence of epidemic threshold in scale-free networks with super-spreaders. Many researchers have been working on analyzing how the dynamics of epidemics are affected by the structure of the complex network being considered[13–15]. The initial conditons of an epidemic play an important role in determining the final size and dynamics of an epidemic[7,16]. Authors in Schumm et al. (2007)[17] analyze the effect of introducing weights for the edges in such a network. They found that the edge-weights help in adding realistic details to network structure. A number of simulaton tools were developed by researchers for prediction of disease epidemics. Two such tools developed recently were EpiSims (Epidemiological Simulation System) from Los Alamos

National Laboratory[18] and STEM (Spatiotemporal Epidemiological Modeler) from IBM[19]. The connectivity of a network also plays an important role in determining the dynamics of the epidemics and previous work related to this aspect helps in modeling of epidemics[20]. Among the network models developed for plant diseases a more recent one was the network model by presented in Margosian et al. 2009[21] where the authors consider a network of US counties with the links between adjacent counties and the effect of crop density in these counties and analyzes the connectivity of the landscape based on different threshold levels of densities. A notable model for the spread of soybean rust is IAMS (Integrated aerobiological modeling system) as described in Isard et al. (2006)[22] and Isard et al. (2007)[23]. This model characterized the various stages involved in the spread of the disease, namely the release of spores into the atmosphere, atmospheric transport, exposure to solar radiation, deposition of the spores, host development and disease growth on the host.

Soybean rust, a fungal disease of soybean, has affected a large number of states in the USA. It was detected in the United States in 2004[24]. The disease has already caused great losses in other countries[25–27]. For example, in Brazil huge losses were reported in 2003[28]. The disease spreads through fungal spores carried by the wind and the weed kudzu can act as a reservoir for the pathogen[29]. The disease overwinters in the southeast and migrates annually to the north of the United States[3].

The objective of this thesis is to present a model for the spread of diseases such as soybean rust using complex networks. The centroid of each county in the US was considered to be a node or vertex, where some counties, especially in the eastern US, contain sentinel plots. It was assumed that the sentinel plot and the area around it behave in a similar manner. Sampling in the sentinel plots is generally done every two weeks from the time that the soybeans are a couple weeks old. Sampling then takes place at most every week when the soybeans reach the susceptible stage. Sentinel plots include early-maturing cultivars that can be infected before commercial cultivars, so that northward movement of the pathogen can be more readily detected.

The model presented in this thesis is a modification of the SI model where each entity is classified as either susceptible or infected. It could be considered a SIR model, but the 'removed' (R) stage occurs only as a function of seasonality of the soybean crop, rather than as a step explicitly following infection. Since we were considering the epidemic at the county scale, we did not include the possibility of recovery for a county.

Infection levels at the county scale may also be influenced by selective removal of some severely infected fields. Also, the use of fungicide slows down the spread of soybean rust to an extent. (Fungicide use is more prevalent in the commercial farms of southern United States. However, sentinel plots do not use any fungicides).

A weighted network is considered in which the weights on the edges are based on the wind speed and direction and the host density, which are the driving factors for the spread of the disease. Also taken into account is the availability of the weed kudzu which acts as a reservoir for the pathogen and results in faster movement of the disease in regions where the weed is abundant. The weights were scaled appropriately in order to maintain the probabilities between 0 and 1. The weight parameters were estimated such that they gave minimum errors over the construction datasets. The optimum weights thus estimated were applied to predict the evolution of the disease over the validation datasets. The rust status data collected from the soybean sentinel plots in the United States formed the basis for the construction and validation of the prediction model.

Collection of infection incidence data is critical for any infection monitoring and prediction process. Extensive collection of such data involves huge cost. Thus it becomes important to collect data efficiently with least possible cost. Establishment and maintenance of infection monitoring sites to collect data requires cost. The cost involved can be reduced by reducing the number of monitoring sites. In this thesis, various strategies to reduce the number of monitoring sites have been discussed in order to reduce the cost involved while still maintaining disease prediction accuracy.

The remaining parts of this thesis have been divided as follows. Chapter 2 describes

the data sources and the nature of the data. Chapter 3 describes in detail the structure of the prediction model. Chapter 4 discusses the strategies to effectively reduce the number of infection monitoring sites. Chapter 5 contains the analysis and discussion of results of the disease prediction model and strategic reduction of monitoring nodes. Finally, chapter 6 concludes the findings and discusses future work.

Chapter 2

Datasets

Any modeling process relies heavily on accurate and complete data for construction as well as validation of the model. Also important is to make the best use of the available data in terms of extrapolating the available data for the entire range of the analysis.

2.1 Nature of the data

One of the first tasks was to secure useful data from various sources. The data from the US network of soybean sentinel plots from the years 2005 to 2008 was used to fit our model. These sentinel plots were established for monitoring and research of soybean rust just after the disease was detected in the United States. The rust dataset for each of the years from 2005 to 2008 was comprised of rust status (whether infection was found or not) for a given sentinel plot and the date of observation. The majority of sampling was done on a weekly to biweekly basis. Though this dataset did not cover all the counties in the United States and also each county was not sampled for each of the time-periods, it was the best dataset available for analysis. The data for the western counties is less frequent when compared to that for eastern counties as shown in the figure 2.1. The infection was mostly concentrated in the Southeastern United States as shown in figure 2.2, one of the reasons being the presence of the weed kudzu in this region. In order to estimate the rust condition in the counties which were missing in the dataset, we followed an algorithm as described later in this chapter.

Figure 2.1: *The figure above shows the soybean rust status for the counties of United states for the year 2007. Red nodes represent counties where infection was observed atleast once during the year, green nodes represent counties where no infection was found during the year, grey counties nodes represent counties where no observation was made during the year.*

The host abundance data also from the years 2005 to 2008 was accessed from the US National Agricultural Statistics Service (http://www.nass.usda.gov/Data_and_Statistics/index.asp). This data comprised of the county and state name and FIPS identification number which is unique for a county and the corresponding Soybean area in acres. This included the number of soybean acres planted in a specific year for a given county. The soybean density was computed by normalizing the soybean acreage with the total county acreage. We also used kudzu abudance data which also was comprised of the county and state name and FIPS identification number and the corresponding kudzu area in acres. The density of kudzu was obtained by normalizing the kudzu acreage with the total county acreage. Wind data from first order weather stations was used from the National Climatic Data Center's website. This data included the daily average resultant wind speed and wind direction for first order weather stations in the United States. First order weather stations are mostly airport based weather stations and they are the only weather stations which measure wind velocity but are very limited in number and coverage. Since all the counties in the US do not have a first order weather station, we used the average wind velocity for each state.

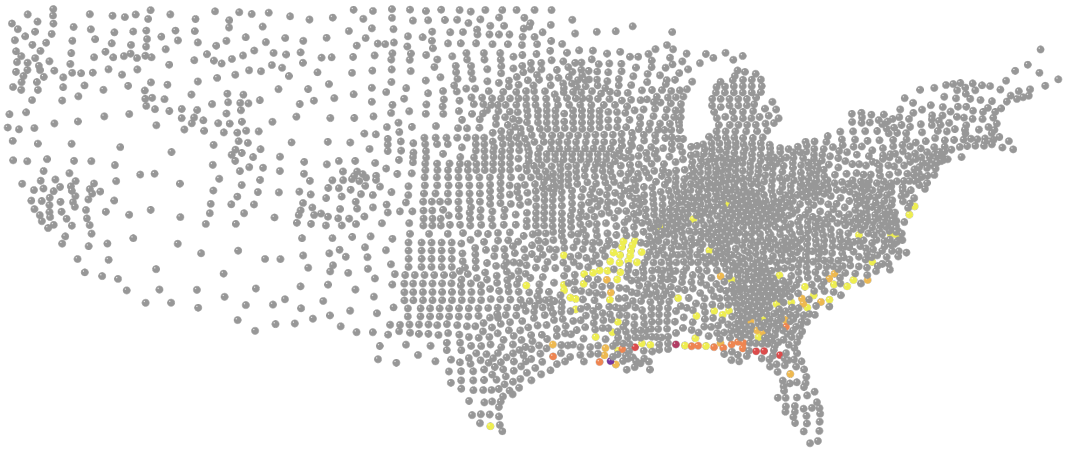

Figure 2.2: *The figure above shows the frequency of infection observed at each node. The colors lilac, dark pink, red, orange, gold, yellow and grey represent infection frequencies above 6, above 4, 4, 3, 2, 1, 0 respectively.*

2.2 Method for estimation of missing spatial data

The Dataset that we used was incomplete as some of the counties did not have rust data for all the time-periods of the year and some counties did not have observations at all, so we estimated the missing data using the following algorithm:

1. Identify the nodes which have never had observations in the current year up to this date.

2. For these nodes, build a list of the nodes which have available data on rust status and which lie within 0.5 degrees of the given node.

3. For those nodes that did not have any neighbors within the defined radius, repeat step 2 by expanding the radius of the circle by 0.25 degrees. Repeat expansions of the radius until the given missing node has at least one node with available data within the radius.

4. After obtaining the list of neighbors for each of the missing nodes, estimate the rust status of the missing nodes by taking the mode of the rust status data of its neighbors.

The values of radii we used for estimating the missing data were 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.00, 2.50, 3.00, 3.50, 4.00, 4.50, 5.00, 10.00, 20.00, 30.00, 40.00, 50.00. The initial radius was chosen such that it covers sufficient number of neighbors within its radius but at the same time does not suffer from data aggregation due to a large number of data points.

Conceptualizing the process of assigning values to missing data is an interesting problem. Some points with missing data may have no or little soybean present. The model will take this into account when movement between any two nodes is estimated. Thus, estimating missing data for a node without soybean is based on determining whether sufficient inoculum for infection is present, not whether infection per se has occurred.

The steps involved in preprocessing of the rust status data are

1. Assemble the information about which nodes are currently observed as infected or uninfected.

2. Also assign as infected any nodes which have been observed to be infected earlier during this year, even if observed as non-infected at this time point.

3. Assign other node values based on the missing value algorithm above.

The steps as described above for estimation of missing data were implemented at each time-step transition before applying the disease prediction model. The next chapter, chapter 3 describes the structure of the model and the variations of the model analyzed.

Chapter 3

Disease Prediction Model

We used the SI model which classifies nodes as being susceptible or infected. It could be considered a SIR model, but the 'removed' (R) stage occurs only as a function of seasonality of the soybean crop, rather than as a step explicitly following infection. We incorporated static as well as dynamic features of the network into the model, the static components being the host density (soybean crop density and kudzu weed density) and the dynamic component being the wind conditions which can be different at each timestep. We modeled the edge-weights as a function of these components.

3.1 Model Description

The key parameters of the model are $\omega_{i,j}$ and $\beta_{i,j}(t)$ which are combined into a single parameter $u_{i,j}(t)$. $u_{i,j}(t)$ is a function of both these parameters and signifies the edge-weight. Here, $\omega_{i,j}$ is a function of the parameters which are taken as constant during the season, which are distance between the nodes, crop density and kudzu density. $\omega_{i,j}$ has a linear relationship with density and decays exponentially with distance. $\beta_{i,j}(t)$ indicates the projection of the scaled monthly average wind velocity onto the direction of the link, which varies with time; it gives the wind-based infection rate between nodes i and j at time t[30]. Seasonal 'removal' of soybeans outside their typical season of production was implemented in the model by removing nodes above 44.00 latitude for the winter months. (In epidemics observed in the US to this point, soybean rust has not approached northern counties prior

Table 3.1: *Symbols used in the model*

Parameters	Definition
$\beta_{i,j}(t)$	Wind based component of infection rate between nodes i and j at time t.
$\omega_{i,j}$	Density and distance based component of infection rate between nodes i and j at time t.
$u_{i,j}$	Edge-weight based on distance, density and wind between nodes i and j at time t.
t	Time step.
$p_{i,t}$	Probability that node i is infected at time t.
$\zeta_{i,t}$	Probability that node i will not receive infection from its neighbors at time t.

to the soybean season.) Over time we update the value of $\beta_{i,j}(t)$ as a function of the wind data, and compute the probability of infection of a node at each time step.

The weight between the nodes i and j, $\omega_{i,j}$ is a function of the parameters which are constant during the season, such as distance between the nodes, soybean density and kudzu density. (In the case of soybean density, there is a constant maximum level attained while the density will be zero at other times of the year, as a function of season.) The soybean density and kudzu density are added together to provide a total host density. The first equation 3.1 represents an exponential model for dispersal and the second equation 3.2 represents a power law model for dispersal. The third equation 3.3 indicates an exponential model where density is incorporated as a product rather than a sum, following a gravity model of density effects[31].

$$\omega_{i,j} = a_1(\frac{d_i + d_j}{2})e^{-a_2 l_{i,j}} \tag{3.1}$$

$$\omega_{i,j} = b_1(\frac{d_i + d_j}{2})l_{i,j}^{-b_2} \tag{3.2}$$

$$\omega_{i,j} = a_1(d_i d_j)e^{-a_2 l_{i,j}} \tag{3.3}$$

Where d_i is the proportion host density (area of soybean or kudzu/total area) in node i and $l_{i,j}$ is the distance between nodes i and j.

The effect of wind can be incorporated in the model by considering infection rate $\beta_{i,j}(t)$ to be proportional to the scalar projection of the wind in the direction of the link between the two nodes i and j (equation 3.5). We used a straight line relationship between the wind-based infection rate i.e. $\beta_{i,j}(t)$ and the wind as shown in the equation 3.4.

$$\beta_{i,j}(t) = comp_{\bar{l_{i,j}}}\bar{w_t} \tag{3.4}$$

$$comp_{\bar{l_{i,j}}}\bar{w_t} = \frac{\bar{l_{i,j}}\bar{w_t}}{|\bar{l_{i,j}}|} \tag{3.5}$$

Where $\bar{l_{i,j}}$ is the distance vector between the two nodes i and j, $\bar{w_t}$ is the wind vector at a time t and $comp_{\bar{l_{i,j}}}\bar{w_t}$ is the scalar projection of the wind vector at time t in the direction of the link between the two nodes, which was normalized by 13mph (a speed determined based on the maximum monthly-average wind-speed observed in any county during any month of the years considered). If the scalar projection of the wind is negative, then it is not considered (it is replaced with zero).

We then combined the $\omega_{i,j}$ and $\beta_{i,j}(t)$ into a single parameter $u_{i,j}(t)$ representing the edge-weight. This gave us more flexibility in analyzing different kinds of interactions among the distance, host density and wind. These interactions have been shown in detail in section 3.2.

The probability $\zeta_{i,t}$ of a node v_i not receiving infections from its neighbors v_j , is expressed as in equation 3.6 in terms of $\omega_{i,j}$ and $\beta_{i,j}(t)$ or as in equation 3.7 in terms of $u_{i,j}(t)$.

$$\zeta_{i,t} = \prod_j (1 - \omega_{i,j}\beta_{i,j}p_{t,j}) \tag{3.6}$$

$$\zeta_{i,t} = \prod_j (1 - u_{i,j}p_{t,j}) \tag{3.7}$$

Where $p_{i,t}$ is the probability of node v_i being infected at time t. Here, the product of $\omega_{i,j}$ and $\beta_{i,j}(t)$ is in [0, 1]. The probability of node i being infected at a time t is expressed in equation 3.8.

$$p_{i,t} = 1 - (1 - p_{i,t-1})\zeta_{i,t} \tag{3.8}$$

3.2 Model variations for the structure of weights

We tested the different kinds of interactions between distance, host density and wind components for characterizing the edge-weights in our network.

3.2.1 Multiplicative model using gravity law

This model is a fully multiplicative model where edge-weights are computed by multiplying all elements, namely density, wind and exponential distance factor, also considering the product of source and destination crop densities to compute the weights based on the gravity law.

$$u_{i,j} = a_1(d_i d_j)\frac{\bar{l_{i,j}}\bar{w_t}}{|\bar{l_{i,j}}|}e^{-a_2 l_{i,j}} \tag{3.9}$$

3.2.2 Multiplicative model with average density

Multiplicative model uses edge-weights that are computed by multiplying all elements, namely density, wind and exponential distance factor. Here, we use the average of the source and destination densities.

$$u_{i,j} = a_1(d_i + d_j)\frac{\bar{l_{i,j}}\bar{w_t}}{|\bar{l_{i,j}}|}e^{-a_2 l_{i,j}} \tag{3.10}$$

3.2.3 Additive model

Additive model uses edge-weights that are computed by performing a weighted sum of all elements, namely density, wind and exponential distance factor.

$$u_{i,j} = a_1(d_i + d_j) + a_2\frac{\bar{l_{i,j}}\bar{w}_t}{|\bar{l_{i,j}}|} + a_3 e^{-a_2 l_{i,j}} \tag{3.11}$$

3.2.4 Hybrid models

The Hybrid models include various combinations of the multiplicative and additive models.

Pair-wise interaction model with density-distance and wind-distance interaction

$$u_{i,j} = [a_1(d_i + d_j) + a_2\frac{\bar{l_{i,j}}\bar{w}_t}{|\bar{l_{i,j}}|}]e^{-a_2 l_{i,j}} \tag{3.12}$$

Additive plus three-way interaction model

$$u_{i,j} = a_1(d_i + d_j) + a_2\frac{\bar{l_{i,j}}\bar{w}_t}{|\bar{l_{i,j}}|} + a_3 e^{-a_2 l_{i,j}} + a_4(d_i d_j)\frac{\bar{l_{i,j}}\bar{w}_t}{|\bar{l_{i,j}}|}e^{-a_2 l_{i,j}} \tag{3.13}$$

Additive plus pair-wise interaction model

$$u_{i,j} = a_1(d_i + d_j) + a_2\frac{\bar{l_{i,j}}\bar{w}_t}{|\bar{l_{i,j}}|} + a_3 e^{-a_2 l_{i,j}} + a_4(d_i + d_j)\frac{\bar{l_{i,j}}\bar{w}_t}{|\bar{l_{i,j}}|} + a_5(d_i + d_j)e^{-a_2 l_{i,j}} + a_6\frac{\bar{l_{i,j}}\bar{w}_t}{|\bar{l_{i,j}}|}e^{-a_2 l_{i,j}} \tag{3.14}$$

For each of the variations of the model described above, the parameters were estimated and performance was evaluated and compared with other models. The summary of the performance of these models has been covered in chapter 5 of this thesis.

3.3 Error calculation and Parameter estimation

For the observed dataset, a value of 1 is assigned to the counties/nodes which are observed as being infected and a value of 0 is assigned to the nodes which are observed not infected. The simulation generates the predicted probabilities for each node being infected. The simulation results are compared with the observed data by taking the absolute value of the difference

between the predicted probabilities and value of the observed data (1 or 0 depending on whether it is observed to be infected or not). The error in prediction for each of the nodes is then mapped. The total error is obtained by summing up the individual errors for all the counties/nodes, with weighting structure discussed below. The parameters in the model were estimated by evaluating model fit for different combinations of the parameters. The set of parameters which gave the least total error were chosen.

The mean error in infected nodes for the time-step t can be computed as

$$E_{in} = \sum_{i=1}^{n_{in}(t)} \frac{1 - p_i(t)}{n_{in}(t)} \tag{3.15}$$

Where i =1, 2, 3...$n_{in}(t)$ and $n_{in}(t)$ is the total number of infected nodes in the given timestep t (month). Similarly, the mean error in healthy nodes for the time-step t can be computed as

$$E_{hn} = \sum_{j=1}^{n_{hn}(t)} \frac{p_j(t)}{n_{hn}(t)} \tag{3.16}$$

Where s=1, 2, 3...$n_{hn}(t)$ and $n_{hn}(t)$ is the total number of healthy nodes in the given timestep(month).

The overall error will be a weighted sum of the mean error in the observed-infected nodes and the mean error in observed-healthy nodes

$$E = \alpha E_{in} + \beta E_{hn} \tag{3.17}$$

where α and β are the weights assigned to the error in observed-infected nodes and error in observed-healthy nodes respectively and $\alpha + \beta = 1$. The observed-infected nodes were given 9 times more weight than the observed-healthy nodes for evaluating the final error i.e. α:β=9:1.

Chapter 4

Strategic reduction of infection monitoring nodes

Sentinel plots play a very important role in the monitoring process of plant diseases in a given area/county of the country. The data collected from these plots are vital for construction and validation of predictive models. Establishment and maintenance of sentinel plots is expensive. Our aim is to minimize the cost involved by sampling the current set of sentinel plots such that the cost of establishment, maintenance and monitoring are minimized while not degrading the prediction accuracy of the model. The different components involved with sentinel plots are

1. Establishing a sentinel plot.

2. Maintenance of the sentinel plot.

3. Sampling frequency for the sentinel plot.

Described below are the different methods for positioning a reduced number of sentinel plots. Since we used existing sentinel plot information in our analysis, the number of sentinel plots sampled in each monthly transition varied slightly. We based the reduction in percentage plots available for each monthly transition on the number of plots with information available for that transition. The following methods of selecting sentinel plots for inclusion represent increasingly sophisticated approaches for including information about

the epidemic. We evaluated the error resulting when x% (10%, 25%, 50%, 75% and 100%) of the original observed sentinel plot network was used for making model predictions for 16 monthly transitions as these had substantial observations and higher infections. For cases where sampling nodes were removed at random, we generated 50 realizations.

4.1 Random selection of informative nodes

The simplest way of reducing the total number of sentinel plots is to randomly sample the entire observed set of sentinel plots. We evaluated the error resulting when x% (10%, 25%, 50%, 75% and 100%) of the original observed sentinel plot network was used for making model predictions. The graph of the error in prediction versus the percentage of original counties included in the sampling network exhibited an exponentially decaying behavior.

4.2 Zonal selection

In this method, we exploit the fact that disease has always been found only in the Southeastern US and has rarely reached the north or the west. This approach follows preferential sampling in the Southeast and reduced sampling elsewhere. Here, we have more number of nodes in the Southeast and fewer nodes in the remaining regions. This way we have a higher density of plots in regions of greater observed frequency of infection. We divided the country into 3 zones as follows:

1. Region1 between 25.61 and 38 degrees latitude and -98 and -67.63 degrees longitude. This is the Southeastern region with highest infection frequency.

2. Region2 between 38 and 44 degrees latitude and -110 and -98 degrees longitude.

3. Region3 between 44 and 48.77 degrees latitude and -124.15 and -110 degrees longitude.

We maintained a density of 80%, 10% and 10% of the total number of informative nodes in the network for region 1, 2 and 3 respectively.

4.3 Infection frequency based selection

In this method, we calculate the frequency with which each node has been observed infected for the entire dataset, including observations outside the 16 active monthly transitions in order to better characterize starting conditions, and then order the nodes from highest to lowest values of frequency. The resulting network consists of nodes with non-zero frequency or frequency above a certain number. The infection frequency of nodes has been illustrated in the figure 2.2.

4.4 Combination of infection frequency and node strength based selection

In this method, infection frequency and node strength were weighted in the ratio 80:20 (after they had been scaled to be between 0 and 1) and the nodes were ordered in decreasing order of this weighted value and only the highest $x\%$ of the whole set of nodes were considered.

Chapter 5 contains the analysis and discussions of the results of disease prediction model and strategic reduction of disease monitoring sites.

Chapter 5

Results and Discussions

The analysis of the results has been divided into results of the disease prediction model and parameter calibration, validation of importance of using host density and wind velocity data in the model, results of analysis of strategic reduction of montoring sites and the discussion of these results.

5.1 Results of Disease Prediction

We applied our model to make predictions for the summer months of the years 2005 to 2008. We chose to focus on the summer months from May to June for the years for which we had data because soybean rust was not active during the other months due to cold winter temperatures which were not suitable for the pathogen to survive and propagate. We used year 2007 data for construction of the model and years 2005, 2006, 2008 for validation of our model. We analyzed two kinds of predictions into the future, one-step predictions and multi-step predictions. For one-step predictions, we used data from a first time-step to predict the next time-step, and the third time-step using the second and so on. This approach is very useful when we have up-to-date data coming in and need predictions into near future. This would help farmers decide about fungicide use depending on the prediction. For multi-step predictions, we use a single time-step to predict many time-steps into the future. This is particularly useful when we need to make predictions for an entire season or year or farther in future even if we do not have data for intermediate steps. This is more important from

Figure 5.1: *Observed rust status in August 2007. Red nodes represent counties where infection was observed at least once during the time period, green nodes represent counties where no infection was found during the time period, grey nodes represent counties where no observation was made during the time period.*

Table 5.1: *Multiplicative model with gravity model for densities: Error percentages for different time steps*

Year	May-June	Jun-Jul	Jul-Aug	Aug-Sep
2005	No infection	2.124	2.468	3.313
2006	2.476	3.545	1.112	2.081
2007	1.060	1.480	2.681	4.388
2008	3.411	2.598	3.139	0

the management perspective and establishment of new plots. Shown in figures 5.1, 5.2, 5.3 are the various steps involved in the prediction process are the results for the prediction from May 2007 to Jun 2007. It shows the maps for observed rust status, the result after estimation of missing data and the prediction for the next time step.

The parameter estimates for the Multiplicative gravity model for all the year 2007 monthly steps over summer months, i.e., May 2007 to June 2007, June 2007 to July 2007, July 2007 to August 2007, August 2007 to September 2007, were $a_1 = 0.01$ and $a_2 = 1$.

The parameter estimates for the Multiplicative model with average density of source and destination for all the year 2007 monthly steps over summer months, i.e., May 2007 to June

Figure 5.2: *Estimation of missing data. Red nodes represent counties which were estimated to be infected, green nodes represent counties which were estimated to be uninfected.*

Figure 5.3: *Prediction for September 2007. Dark red nodes represent counties which were predicted to be infected with high probability, green nodes represent counties which were predicted to be uninfected with negligble probability of infection, all other shades from green to dark red represent increasing probabilties of infection.*

Table 5.2: *Multiplicative model with sum of densities: Error percentages for different time steps*

Year	May-June	Jun-Jul	Jul-Aug	Aug-Sep
2005	No infection	2.872	2.451	4.270
2006	3.978	4.495	3.655	5.023
2007	10.025	14.594	3.043	4.035
2008	3.263	2.049	3.064	1.190

Table 5.3: *Additive model: Error percentages for different time steps*

Year	May-June	Jun-Jul	Jul-Aug	Aug-Sep
2005	No infection	6.485	8.9157	4.0357
2006	4.157	5.800	3.792	3.8491
2007	5.970	9.141	6.959	6.934
2008	3.957	4.421	3.655	5.678

2007, June 2007 to July 2007, July 2007 to August 2007, August 2007 to September 2007, were $a_1 = 0.01$ and $a_2 = 10$.

The parameters estimates for the Additive model with average density of source and destination for all the year 2007 monthly steps over summer months i.e., May 2007 to June 2007, June 2007 to July 2007, July 2007 to August 2007, August 2007 to September 2007 were $a_1 = 0.01$ and $a_2 = 1$ and $a_3 = 2$ and $a_4 = 2$.

5.2 Validation of importance of using host density and wind velocity data in the model

In order to validate the importance of host density and wind, we applied bootstrapping i.e. we randomly sampled the densities and wind values with replacement for all the counties. First, the set of observed county host density values were randomly reassigned with replacement, maintaining the original observed wind speeds and directions for each county. The host densities were randomly reassigned in 500 independent simulations, and the error associated with predictions (based on the parameter estimates for the observed densities) for each simulation was recorded. The observed error was compared to the distribution of

errors from the simulations. The errors from the simulations were ranked from smallest to largest and the position of the observed error within the list was noted. The rank of the observed error among the 500 values of the bootstrap distribution from randomizing host density was found to be '1' for all time-periods which implies that the original host density data gives the least error. All other reassignments of host densities degrade the performance. This clearly shows the importance of using host density data for predictions. Second, the set of wind speeds and directions was randomly reassigned with replacement, maintaining the original observed host densities. The observed error was compared to the distribution of errors from 500 simulations based on the randomly reassigned wind data. The rank of the observed error among the 500 values of the bootstrap distribution from randomizing wind velocity was found to be '1' for all time-periods which implies that the original wind velocity data gives the least error. All other reassignments of wind velocities degrade the performance. This clearly shows the importance of using wind data for predictions.

5.3 Results for Strategic reduction of informative nodes or sentinel plots

We analyzed the effects of random sampling approach on the error in prediction using our model for reduction to 10%, 25%, 50%, 75% and 100% of the total set and plotted the box-plot for 50 runs. The graph shows an exponentially decaying behavior with increase in the percentage of counties considered for random sampling approach. With strategic zonal sampling, a marked improvement in the performance is achieved as shown in figures 5.4, 5.5, 5.6 and 5.7.

The results of random selection of monitoring nodes and zonal selection strategy for the complete dataset from year 2005 to 2008 have been summarized in the figure 5.8. The summary of random selection of monitoring nodes and zonal selection strategy was constructed using the average percentage errors over 50 runs at each timestep from 2005 to 2008 for both strategies. As in the case of 2007 months, the error in the other years is also relatively

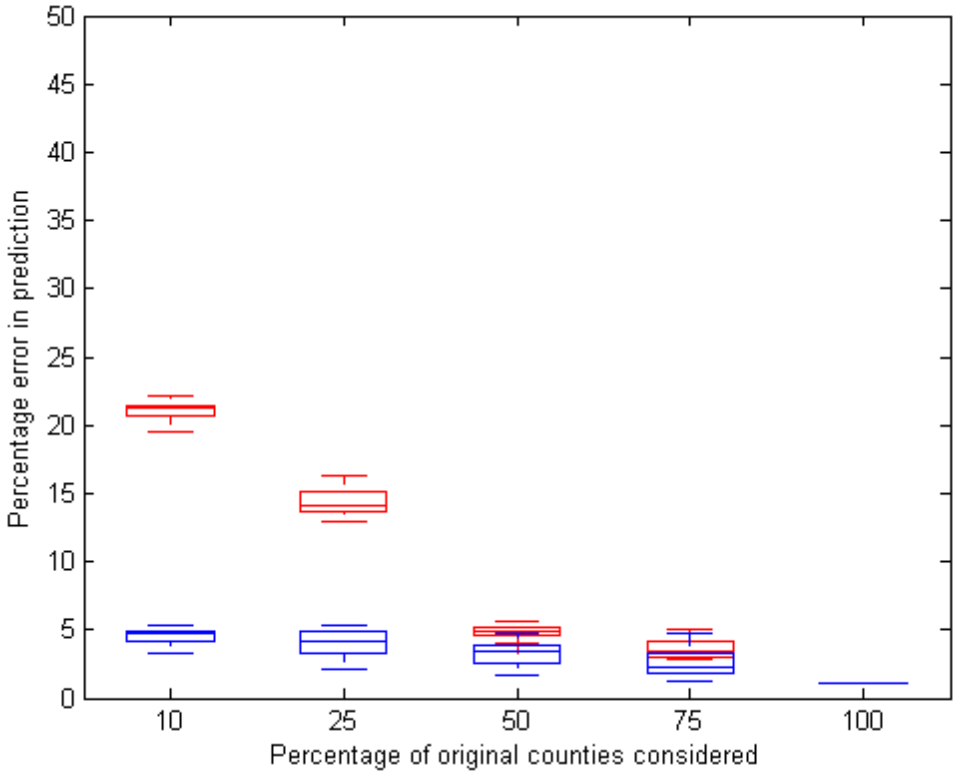

Figure 5.4: *Results of random sampling for May 2007 to June 2007. Red plot indicates results of Random selection, blue plot indicates results of Zonal selection. Strategic Zonal selection gives lower errors when compared with random selection.*

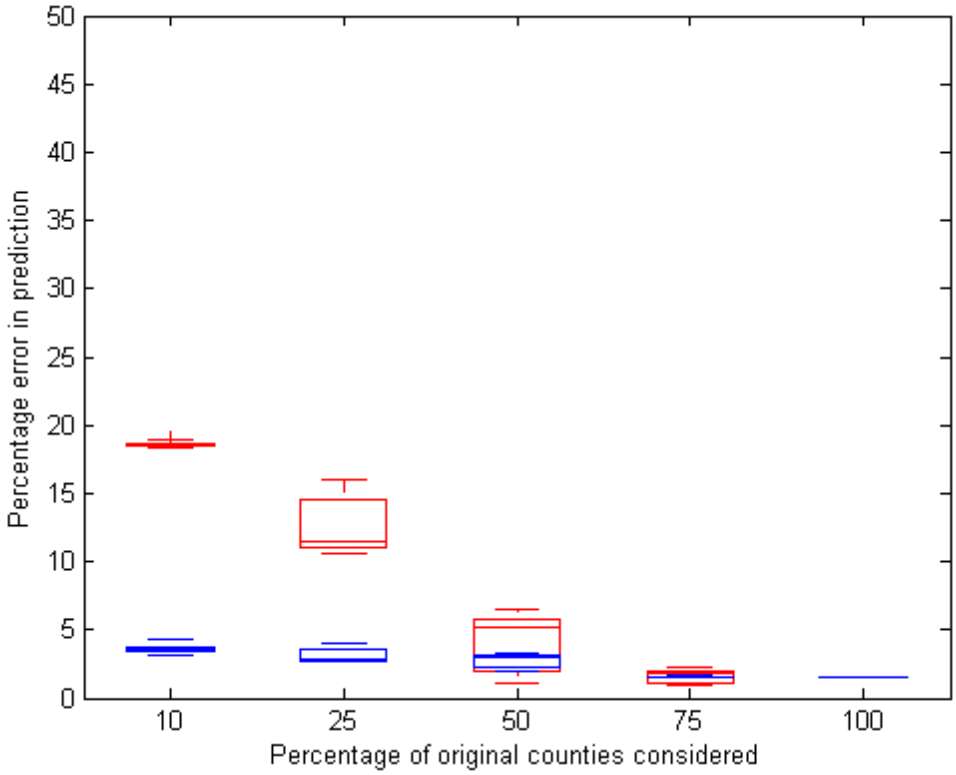

Figure 5.5: *Results of random sampling for June 2007 to July 2007. Red plot indicates results of Random selection, blue plot indicates results of Zonal selection. Strategic Zonal selection gives lower errors when compared with random selection.*

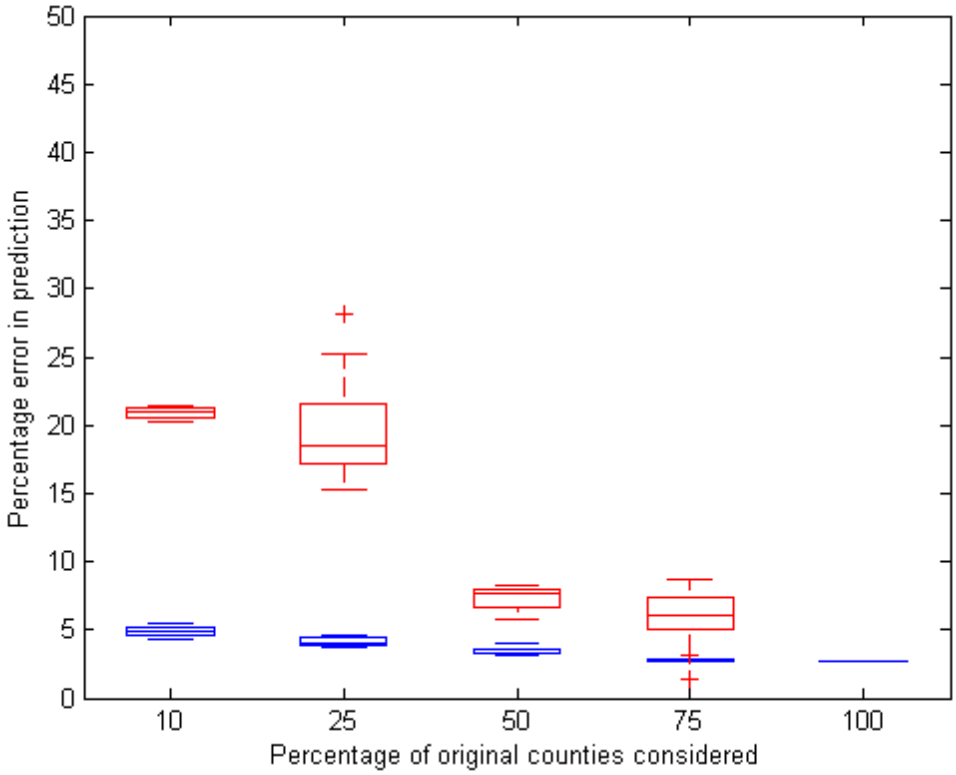

Figure 5.6: *Results of random sampling for July 2007 to August 2007. Red plot indicates results of Random selection, blue plot indicates results of Zonal selection. Strategic Zonal selection gives lower errors when compared with random selection.*

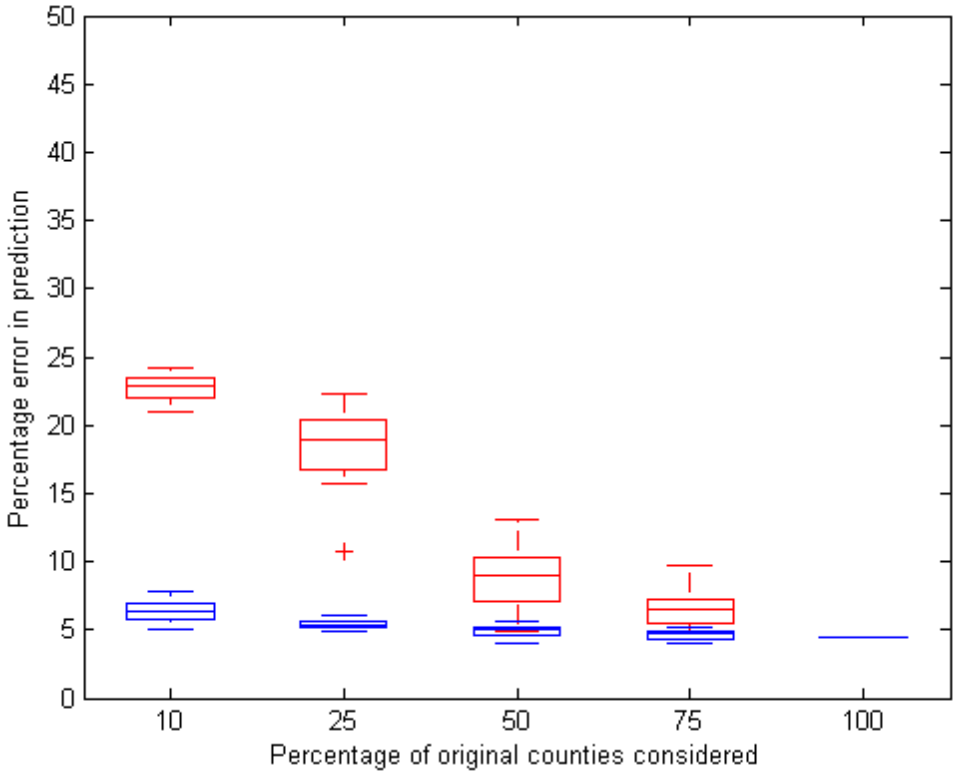

Figure 5.7: *Results of random sampling for August 2007 to September 2007. Red plot indicates results of Random selection, blue plot indicates results of Zonal selection. Strategic Zonal selection gives lower errors when compared with random selection.*

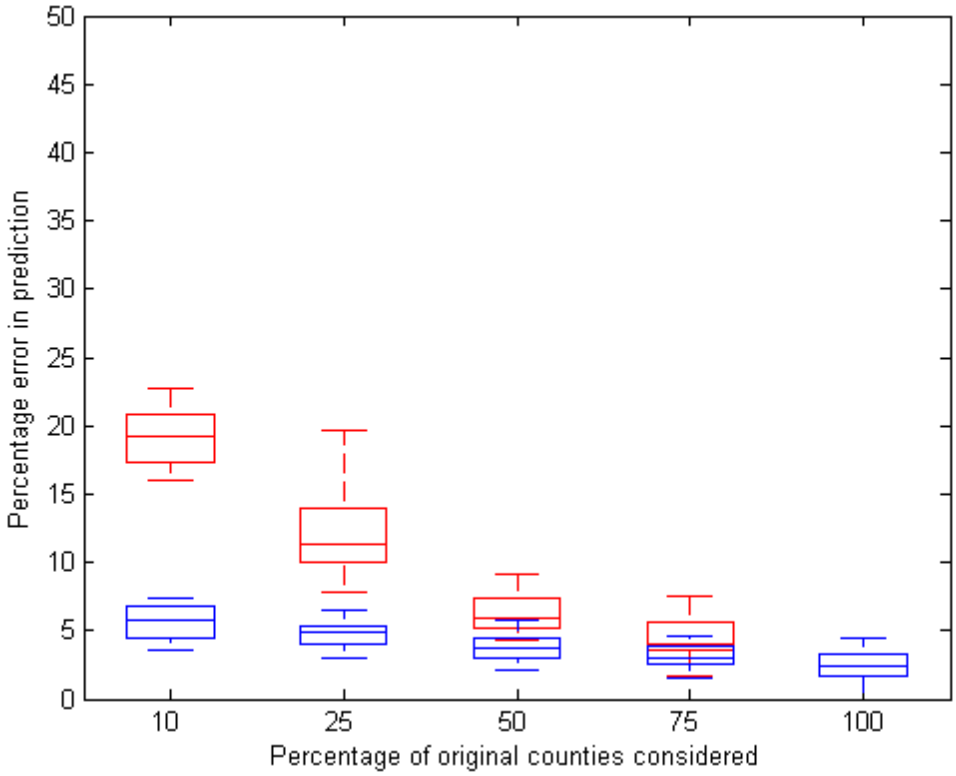

Figure 5.8: *Summary of random selection and zonal selection over all the years. Red plot indicates results of Random selection, blue plot indicates results of Zonal selection. Strategic Zonal selection gives lower errors when compared with random selection.*

higher for random selection process. This error in prediction is reduced significantly using the zonal selection strategy.

The following figures 5.9, 5.10, 5.11, 5.12 show the Comparison of Infection frequency based and combined Infection frequency and nodestrength based selection for the various timeperiods. It can be observed from the plots that addition of nodestrength information to the infection frequency information gives better performance in terms of error in prediction. The total number of nodes at each timestep is about 400 to 600 in number, hence 10% of nodes implies selecting about 40 to 60 nodes.

The results of selection strategies based on node properties, i.e selection based on infection frequency of the nodes and selection based on a weighted sum of infection frequency and

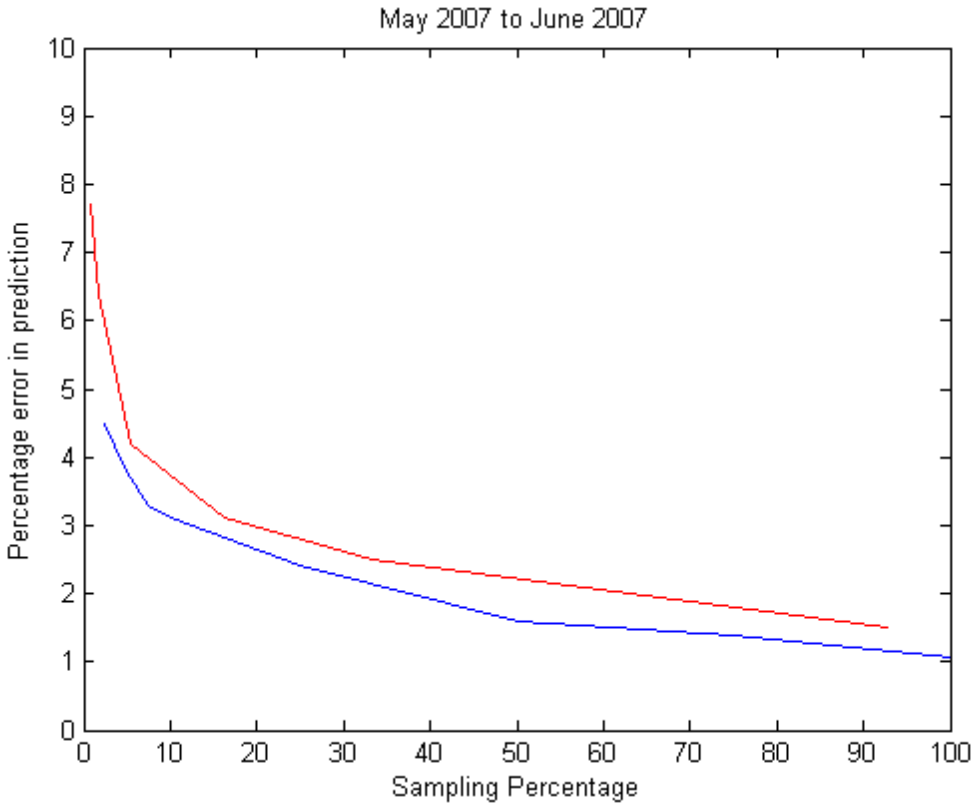

Figure 5.9: *Comparison of Infection frequency based and combined Infection frequency and nodestrength based selection for May 2007 to June 2007 (Red plot for infection frequency based selection and blue plot for combined Infection frequency and nodestrength based selection).*

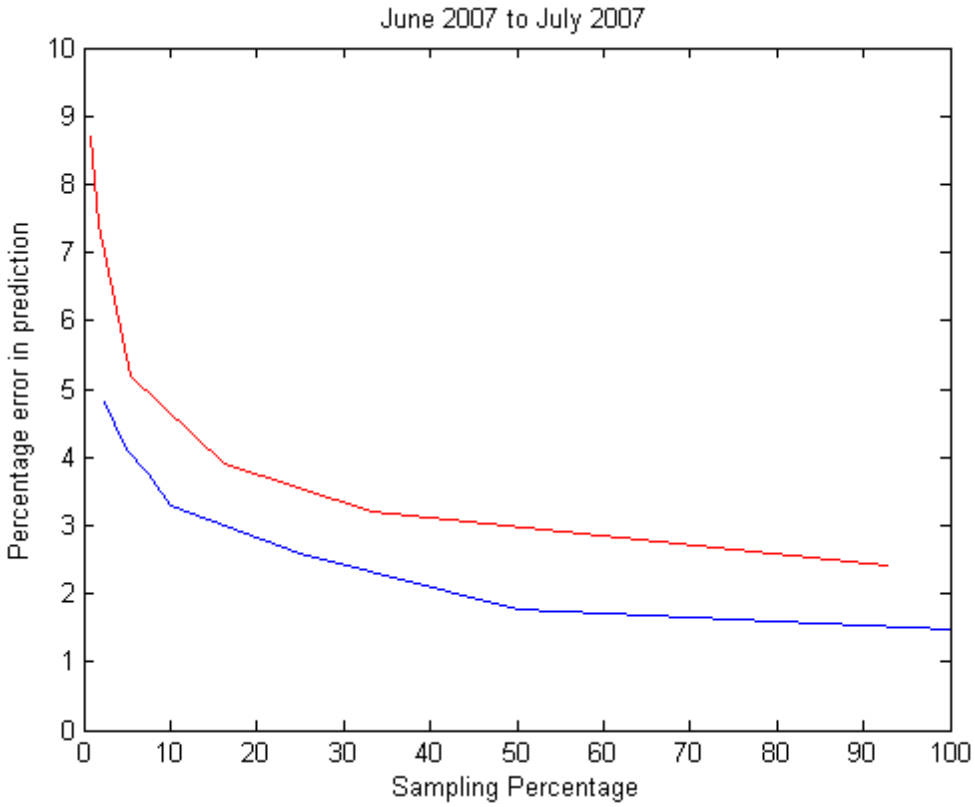

Figure 5.10: *Comparison of Infection frequency based and combined Infection frequency and nodestrength based selection for June 2007 to July 2007 (Red plot for infection frequency based selection and blue plot for combined Infection frequency and nodestrength based selection).*

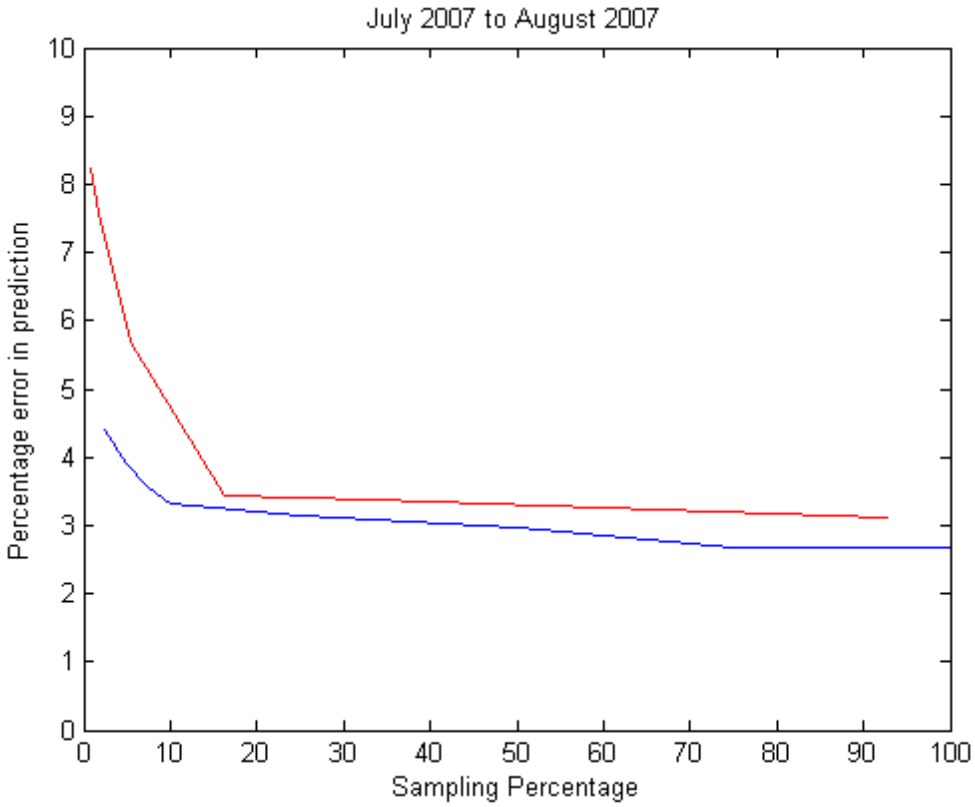

Figure 5.11: *Comparison of Infection frequency based and combined Infection frequency and nodestrength based selection for July 2007 to August 2007 (Red plot for infection frequency based selection and blue plot for combined Infection frequency and nodestrength based selection).*

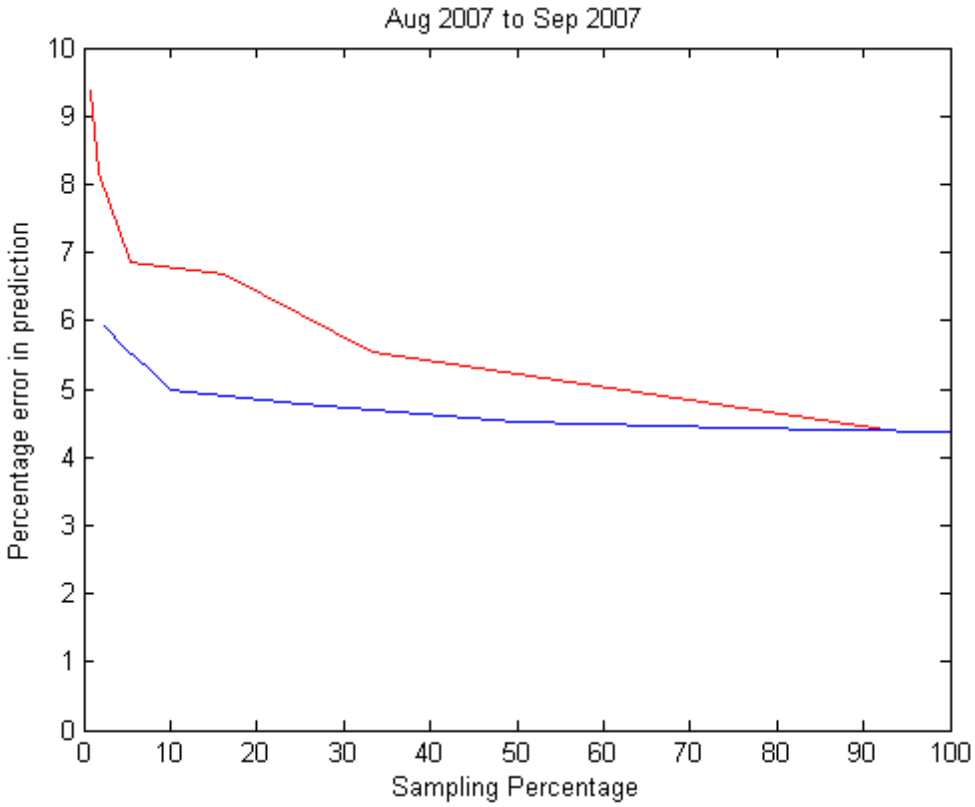

Figure 5.12: *Comparison of Infection frequency based and combined Infection frequency and nodestrength based selection for August 2007 to September 2007 (Red plot for infection frequency based selection and blue plot for combined Infection frequency and nodestrength based selection).*

node strength (weighted in the ratio 80:20 respectively) for the complete dataset from year 2005 to 2008 have been summarized in the figure 5.13. The summary of random selection of monitoring nodes and zonal selection strategy was constructed using the average percentage errors from all timesteps from 2005 to 2008 for both strategies. For the infection frequency based selection, the frequencies varied from zero to ten. Since this strategy could not be evaluated for the exact same percentage of original nodes as the other strategy, the values were plotted by taking the percentage nearest but lower to x-axis mark-up percentages. The node property based selection strategies outperformed the random and zonal selection strategies. The addition of node strength information to the infection frequency information improved the performance significantly.

5.4 Discussion

Among all the models analyzed, some forms of the model seem to perform better over some time-periods than other model variations. Overall, the multiplicative model with densities multiplied as specified by the gravity law[31,32] seemed to do well over a wide range of time-periods. The multiplicative model with the crop densities of the two counties multiplied is mechanistically more sound for this application. But it does not allow analysis of the importance of the different factors involved as we cannot change the weights associated with the distance, density and wind independently. On the other hand, while the additive model allows for weighting the various factors differently, it does not consider the interaction between these factors in its simplest form. The Hybrid models to some extent allow for varying the weights associated with the above mentioned factors and also to have one or more interaction components between these factors.

While the Wang model[10] considers the network to be homogeneous, our model also takes into account the weights associated with the edges[17] and we analyze different structures for these weights in terms of the performance. The network model developed by Margosian et al. (2009)[21] considers the links between adjacent counties and the effect of crop density

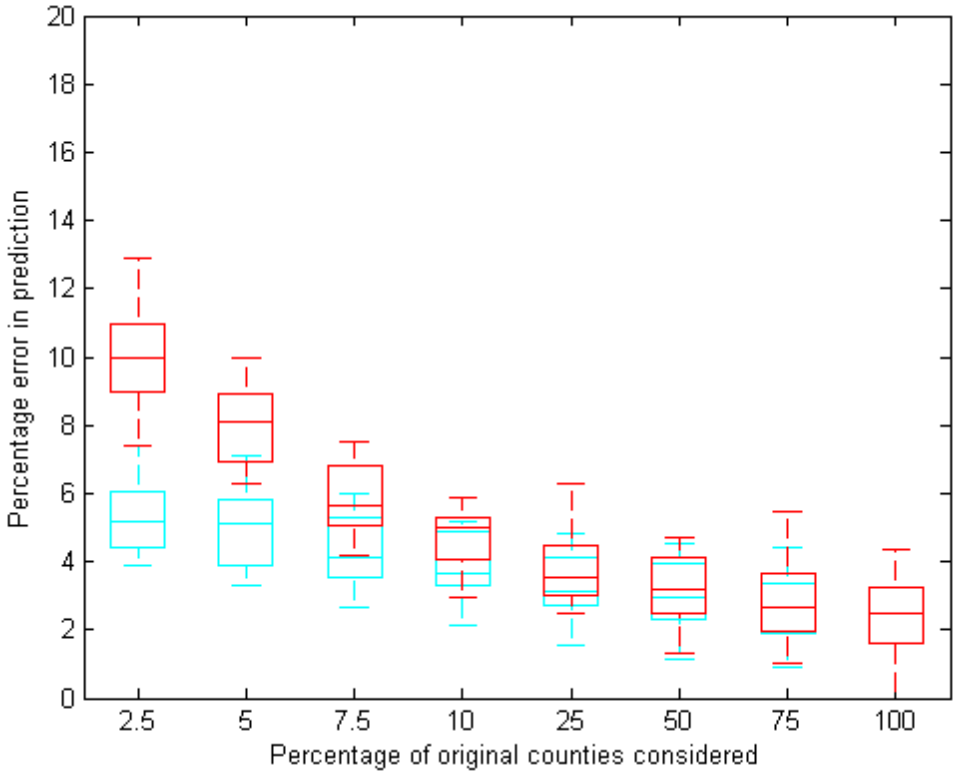

Figure 5.13: *Summary of infection frequency based selection and weighted infection frequency and node strength based selection (weighted in the ratio 80:20) over all the years. Red plot indicates results of infection frequency based selection, blue plot indicates results of weighted infection frequency and node strength based selection. Addition of node strength information to the infection frequencies of the nodes lowers the errors significantly.*

in these counties. Our model further enhances the network by not only considering the adjacent counties but also those that are further apart based on flexible threshold on the distance. Additionally, we also incorporated the effect of wind speed and direction and the availability of kudzu as reservoir for the pathogen.

Application of this model for predicting soybean rust in North America by using latest data from the sentinel plots could prove very beneficial for soybean growers and management, the government and economy as a whole. Our results will have immediate application for soybean rust management and general applications for other plant or animal epidemics or insect infestations studied across large or small landscapes[33]. Such spatio-temporal predictions could aid the growers in optimally timing the planting of soybean and optimal usage of fungicides, thus reducing economic losses. This type of network based approach could be extended to other wind-borne plant diseases and can also be adapted to predict wind-borne diseases in animals and humans. The primary difference in the system for animals and humans would be the mobility which is very negligible in the case of plants. Plants are seldom carried over large distances during the growing season, except for some horticulatural crops. Disease prediction models based on a dynamic network such as the one described here play a critical role in determining the future path of an epidemic and help prepare for outbreaks. Timely predictions complement the mitigation strategies used to control the disease thus lessening the impact of an epidemic.

While our model is sufficiently good at spatio-temporal predictions, there are some areas where our model performance can be enhanced further by making some more improvements. The time-step considered for the predictions is a critical factor in determining the rate of infection. Further analysis at different time-scale resolutions would lead to better insights into the dynamics of the disease. While the model currently takes into account the effect of wind in carrying the spores from one location to another, other factors like temperature[34,35], moisture[36], and UV radiation[37] could also be incorporated. Also incorporating spore trap data in the model as a measure of amount of inoculum or infection present in an area could

further improve the predictions.

Apart from making predictions about outbreaks and their severity in different areas, this model is also being extended for reducing the number of the monitoring sites (sentinel plots) so that predictions can be made with least possible sampling effort and feasible cost. From the analysis of random selection of monitoring nodes, we observed a exponentially decaying behavior for the error in fit with decreasing number of nodes being considered for random selection process. On the other hand, the strategic reduction with preferential selection showed marked improvement even while including only a fraction of the informative nodes. The infection frequency based sampling showed some improvement over zonal sampling strategy. The combination of infection frequency and node-strength information further enhanced the performance of the model allowing us to reduce the number of infection monitoring sites to lower than that achieved by other reduction methods.

Chapter 6

Conclusions

This thesis presented a novel network based model for wind-borne diseases which predicts the spread of the disease with time on a continental scale. The model has been tested for soybean rust which is wind-borne disease of the soybean plant. Inclusion of the effect of host abundance and wind velocity into the model is an important contribution towards improved epidemic prediction models. The model performs well with high accuracies over all the time-periods considered. Among all the variations of the weights considered, the multiplicative model with gravity law, which considers the product of host densities to characterize the weights, performs the best.

From the analysis of reduction of number of disease monitoring sites, which are the soybean sentinel plots in this case, it was observed that it is not required to monitor every node in the network. The monitoring effort should be focussed more on the parts of the network where infection frquency in the past has been high. This would considerably reduce the number of monitoring sites required and thereby reducing the cost involved. Selection process based on observed infection frequency of each node proves to be more beneficial than Zonal sampling (random sampling in regions of higher infection). The infection frequency information of the nodes when combined with the node-strength information gives the best possible reduction in number of monitoring nodes in the network of sentinel plots among the different methods analyzed.

The software tool developed to simulate the prediction model and analyze strategies for

reduction of number of disease monitoring sites can be adapted to other scenarios involving other diseases and/or other base networks.

The disease prediction model can be easily applied to other wind-borne diseases with minor disease specific adaptations or modifications. The findings of the analysis of reduction of number of disease monitoring sites can also be applied for any monitoring network in general.

Future work would involve testing the model with other diseases for varied spatial and temporal resolutions and also incorporating other environmental factors such temparature, precipitaton and solar radiation for characterizing the edge-weights. The model can be improved further by using the observed disease severity within a node when such data becomes available. Future work for strategic positioning of monitoring sites would involve sampling based on other node characteristics like betweenness and clustering coefficient.

Bibliography

[1] X. B. Yang, W. M. Dowler, and A. T. Tschanz, Journal of Phytopathology **133**, 187 (1991).

[2] L. J. Ford, J. Kaufman, and I. Eiron, In: 2nd Natl Soybean Rust Symp. American Phytopathology Society (2006).

[3] M. Livingston et al., Electronic Outlook Report from the U.S. Department of Agriculture Economic Research Service, OCS-04D-02 (2004).

[4] M. J. Roberts, D. Schimmelpfennig, E. Ashley, and M. Livingston, United States Department of Agriculture, Economic Research Service, Economic Research Report No. 18 (2006).

[5] D. L. Urban, E. S. Minor, E. A. Treml, and R. S. Schick, Ecology Letters **12**, 260273 (2009).

[6] M. J. Jeger, Agricultural and Forest Meteorology **97**, 331349 (1999).

[7] M. J. Jeger, M. Pautasso, O. Holdenrieder, and M. W. Shaw, New Phytologist **174**, 279 (2007).

[8] A. Lamour, A. J. Termorshuizen, and D. V. et al., Fems Microbiology Ecology **62**, 222 (2007).

[9] M. Pautasso and M. J. Jeger, Ecological Complexity **5**, 1 (2008).

[10] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, In Proceedings of the Symposium on Reliable Distributed Computing , 25 (2003).

[11] A. V. R. Pastor-Satorras, Phys. Rev. Lett. **86**, 32003203 (2001).

[12] A. V. R. Pastor-Satorras, Phys. Rev. **E 65**, 035108 (2002).

[13] M. E. J. Newman, Phys. Rev. **E 66**, 016128 (2002).

[14] P. V. Mieghem, J. S. Omic, and R. E. Kooij, IEEE/ACM Transaction on Networking **17**, 1 (2009).

[15] M. J. Keeling and K. T. D. Eames, Interface **2**, 295307 (2005).

[16] K. A. J. White and C. A. Gilligan, Journal of Theoretical Biology **242**, 670682 (2006).

[17] P. Schumm, C. Scoglio, D. Gruenbacher, and T. Easton, in Proceedings of IEEE/ACM Bionetics (2007).

[18] S. Eubank, In Proceedings of the 2002 ACM symposium on Applied computing (SAC '02) , 139 (2002).

[19] D. Ford, J. Kaufman, and I. Eiron, International Journal of Health Geographics **5**, 4 (2006).

[20] R. Albert and A. L. Barabasi, Rev. Mod. Phys. **74**, 4797 (2002).

[21] M. L. Margosian, K. A. Garrett, J. M. S. Hutchinson, and K. A. With, BioScience (2009).

[22] S. A. Isard and J. M. Russo, In: 2nd Natl Soybean Rust Symp. American Phytopathology Society (2006).

[23] S. A. Isard, J. M. Russo, and A. Ariatti, Aerobiologia **23**, 271 (2007).

[24] W. Schneider, C. A. Hollier, H. K. Whitman, M. E. Palm, and J. M. McKemy, Plant Disease **89**, 774 (2005).

[25] K. R. Bromfield, American Phytopathological Society (1984).

[26] J. B. Sinclair and G. L. Hartman, National Soybean Research Laboratory Publication Number 1 (1996).

[27] X. B. Yang, W. M. Dowler, A. T. Tschanz, and T. C. Wang, Journal of Phytopathology **136**, 46 (1992).

[28] J. T. Yorinori et al., Plant Disease **89**, 675 (2005).

[29] T. N. Lynch, M. R. Miles, and R. D. Frederick‘, In: 2nd Natl Soybean Rust Symp. American Phytopathology Society, St. Louis, MO (2006).

[30] S. A. Isard, S. H. Gage, and P. Comtois, BioScience **55**, 851 (2005).

[31] N. Masuda, H. Mida, and N. Konno, Physical Review **71** (2005).

[32] Y. C. Xia, O. N. Bjornstad, and B. T. Grenfell, American Naturalist **164**, 267 (2004).

[33] M. E. Irwin, Agricultural and Forest Meteorology **97**, 235248 (1999).

[34] J. K. Kochman, Australian Journal of Agricultural Research **30**, 273 (1979).

[35] L. E. Sconyers et al., vailable via APSnet (http://www.apsnet.org/online/feature/sbr/) (2006).

[36] M. A. Marchetti, J. S. Melching, and K. R. Bromfield, Phytopathology **66**, 461 (1976).

[37] S. A. Isard et al., Plant Disease **90**, 941 (2006).

Appendix A

GNU Free Documentation License

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a

printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "**Document**", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "**you**". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "**Modified Version**" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "**Secondary Section**" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "**Invariant Sections**" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "**Cover Texts**" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this

License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "**Transparent**" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "**Opaque**".

Examples of formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "**Title Page**" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "**publisher**" means any person or entity that distributes copies of the Document to the public.

A section "**Entitled XYZ**" means a named subunit of the Document whose title ei-

ther is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "**Acknowledgements**", "**Dedications**", "**Endorsements**", or "**History**".) To "**Preserve the Title**" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front

cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be

listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer

review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled “History” in the various original documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of

some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subse-

quently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with ... Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.