EPIDEMICS ON COMPLEX NETWORKS

by

MOHAMMAD REZA SANATKAR

B.S. , Ferdowsi University of Mashhad, Iran, 2006
M.S. , Amirkabir University of Technology, Iran, 2009

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
2012

Approved by:
Co-Major Professor
Karen Garrett

Approved by:
Co-Major Professor
Bala Natarajan

Approved by:
Co-Major Professor
Caterina Scoglio

Copyright

Abstract

In this thesis, we propose a statistical model to predict disease dispersal in dynamic networks. We model the process of disease spreading using discrete time Markov chain. In this case, the vector of probability of infection is the state vector and every element of the state vector is a continuous variable between zero and one. In discrete time Markov chains, state probability vectors in each time step depends on state probability vector in the previous time step and one step transition probability matrix. The transition probability matrix can be time variant or time invariant. If this matrix's elements are functions of elements of vector state probability in previous step, the corresponding Markov chain is non linear dynamical system. However, if those elements are independent of vector state probability, the corresponding Markov chain is a linear dynamical system.

We especially focus on the dispersal of soybean rust. In our problem, we have a network of US counties and we aim at predicting that which counties are more likely to get infected by soybean rust during a year based on observations of soybean rust up to that time as well as corresponding observations to previous years. Other data such as soybean and kudzu densities in each county, daily wind data, and distance between counties helps us to build the model.

The rapid growth in the number of Internet users in recent years has led malware generators to exploit this potential to attack computer users around the word. Internet users are frequent targets of malicious software every day. The ability of malware to exploit the infrastructures of networks for propagation determines how detrimental they can be to the network's security. Malicious software can make large outbreaks if they are able to exploit the structure of the Internet and interactions between users to propagate.

Epidemics typically start with some initial infected nodes. Infected nodes can cause their

healthy neighbors to become infected with some probability. With time and in some cases with external intervention, infected nodes can be cured and go back to a healthy state. The study of epidemic dispersals on networks aims at explaining how epidemics evolve and spread in networks. One of the most interesting questions regarding an epidemic spread in a network is whether the epidemic dies out or results in a massive outbreak. Epidemic threshold is a parameter that addresses this question by considering both the network topology and epidemic strength.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In epidemic networks, nodes can get infected through their infected neighbors. If a healthy node does not have any infected neighbors, it will be healthy forever. SIR (Susceptible-Infected-Removed) and SIS (Susceptible-Infected-Susceptible) are two primary epidemic models in the epidemic literature. In the SIR model, an infected node is not susceptible to malware anymore once it has been cured. In the SIS model, an infected node's state after being cured is susceptible again. Epidemic typically starts with some initial infected nodes and either spreads or dies out according to various parameters such as epidemic strength and topological properties of the network. Epidemic strength is characterized by cure rate and infection rate in continuous time and cure probability and infection probability in discrete time. Epidemic is modeled by Markovian process. Indeed, Markovian epidemic model is a dynamical system whose states are infection probabilities of nodes. It makes us able to use mathematical results in control theory area to study the dispersal of epidemic in networks. Epidemic threshold represents the robustness of networks against epidemic. Larger epidemic threshold shows that epidemic must be more strength to be able to spread out in the network and if epidemic strength is less than epidemic threshold, epidemic dies out.

In chapter 2, we study epidemic evolution on dynamic networks and model the dispersal of disease by discrete time Markov chains. Particularly, we focus on the dispersal of soybean rust in the network of counties of USA. In this case, the network's nodes represent the counties. The link weights which are actually the conditional probability of infection are determined by soybean and kudzu densities, the distances of counties to each other, and wind direction and speed. Except wind direction and speed, other parameters of the network are constant over time. Therefore, involvement of wind in the network makes it to be dynamic and changes at each time step. This model aims at determining how much it is likely that each county gets infected in the future. The counties' infection probabilities are the output of the model. We choose biweekly interval as the time step to run our model.
Soybean rust always starts from the counties in southern states such as Texas and Florida and wind from south to north carries soybean rust spores to northern states. At each time step, the weather data are feeded to the network model and the dispersal of disease is predicted for the next time step. Also, we exploit the observation data up to the current time step to increase the accuracy of prediction. On the other hand, the available observations

are limited to a few number of counties and also sparse in time. Because the observations for all counties and all time steps are required to construct the network, we need to estimate missing observations. One spatial-temporal moving average method has been proposed to estimate the missing observations. The other matter in building the network is to estimate the link weights' parameters. We consider one parameter in the link weight and estimate it based on maximum likelihood estimator using the available observations of the rest of years. The estimated parameters are district for each link and functions of time. It is remarkable that for each year prediction, there are different possible combinations of other years to be chosen to estimate the parameter. Simulation results show that increasing the number of years' data to build the network results in decrease in miss error and increase in false alarm error.

In chapter 3, we study the specific type of epidemic model that cure rate at each node depends on its neighbours' states. Moreover, we analyze the advantage of taking into account the spatial correlation of nodes' states for the accuracy of the model's output. Most of the recent analytical models for epidemic assume the independence between nodes' states, which our simulation results show that it causes significant increase in prediction error especially in transient state of epidemic.

We choose application of malware propagation for our proposed analytical model. The rapid growth in the number of Internet users in recent years has led malware generators to exploit this potential to attack computer users around the word. Internet users are frequent targets of malicious software every day. Malicious software can make large outbreaks exploiting the weakness points of structure of networks. In general, malware can be categorized in two main groups: random and topological scanning ones. Random scanning viruses use random scanning of IP addresses to find the potential targets, while topological worms extract the social information of users from the social services like email contact lists and online social networking sites such as facebook and twitter to find the potential targets. We can differentiate between various malware and their methods to find victims by choosing appropriate topologies as their underlying graphs. For random scanning types of malware, the network topologies can be modeled by Erdos-Renyi (ER) graphs and for topological malware, underlying topologies are specific overlay topologies created by applications.

First, the full model which takes into account the exact spatial dependence between neighbors' states without any approximation is developed. We derive two other models (1) independent model (2) Markov model, which aim at approximation of the correlation function because computing the exact correlation function is computationally too expensive. In the independent model, spatial independence between neighbors' states is assumed to derive the model. However, the detailed network topology and the temporal dependence are considered to develop the model. In the Markov model, we take into account the partial spatial dependence between neighbors' states when computing the two-node joint probability to calculate the conditional infection probability of each node. At the end to examine the importance of taking into account the spatial dependence between nodes' states, we compare the outputs of the markov and independent model with the Monte Carlo simulation results. This comparison is done for different real and synthesized topologies. For all topologies, simulation results show that the Markov model outperforms the independent model. The

Markov model results in far better performance than the independent model in the transient state in most cases. Both models' results closely overlap with the simulation in the steady state. Moreover, it is shown that the gain of employing the Markov model instead of the independent model is more significant if the initial fraction of infected nodes is small. The Markov model yields far better results than the independent model when networks have a low level of connection. Also, we show that the advantage of exploiting the Markov model rather than the independent model is more substantial if epidemic strength is low.

In chapter 4, we propose a novel epidemic threshold for dynamic networks. Answering the question whether epidemic dies out or spreads garnered so much attention in the field of epidemic study. Epidemic threshold is a parameter that addresses this question by considering both the network topology and epidemic strength. Most of the recent published analytical results of epidemic threshold are for static networks. There are a few papers that consider dynamic networks and use the different framework for their mathematical derivations from our work.

It is remarkable that in the most real cases of epidemic spreading, the underlying networks are dynamic. If you consider malware propagation in mobile adhoc networks, the movement of users causes that the adjacency matrix of users Bluetooth connections becomes dynamic. Another example is human disease epidemic. Contact networks between people are not fixed and change over time because people continuously move from one location to another location. Moreover, the underlying networks of animal and plant disease is dynamic because the factors that influence the spread of disease are typically dynamic.

In this chapter, we consider the SIS model for epidemic spread. The dynamic behavior of networks is modeled by randomly choosing the adjacency matrix from sets of matrices. This type of dynamic networks is called switching dynamic networks. First, an epidemic model based on the assumption of independence among the state of nodes is developed. This model is a nonlinear dynamical system and its states are infection probabilities of nodes. Then, we prove that the origin is always one equilibrium point of this time-varying dynamical system and its stability depends on the network topology and the values of cure and infection probability. After that, we linearize the epidemic system equations to determine whether the origin is asymptotically stable or not. It has been shown that if the origin is a stable equilibrium of the system, the epidemic dies out, otherwise it spreads out. Then, the joint spectral radius of a set of matrices is defined to derive the analytical epidemic threshold for dynamic networks. We show that for undirected networks there is simplified version of epidemic threshold since the epidemic threshold for undirected networks depends only on the largest spectral radius of a set of system matrices, evaluation of the epidemic threshold is computationally less expensive compare to directed networks. Moreover, It is proved that the derived epidemic threshold confirms the conventional analytical results for static networks. After that, we derive epidemic threshold for periodic networks. Also, epidemic spread in dynamic regular networks is studied and shown that the epidemic threshold for dynamic regular networks is the same as static regular networks. We derive an upper bound for the probability of an epidemic spreading out in dynamic Gilbert networks. Finally for validation our analytical results, we simulate epidemic in Watts-Strogatz, Barabasi-Albert, Regular, and dynamic Gilber . In addition, we consider epidemic on real networks and do simulation on the MIT reality mining graphs.

Chapter 2

Markovian Model for Disease Prediction

2.1 System Model

2.1.1 Introduction

In this chapter, we propose a new method based on maximum likelihood estimation to build a statistical model to predict disease dispersal in dynamic networks. We model the process of disease spreading using discrete time Markov chain. The links weights represent the conditional probability of infection of the destination node given the source node is infected. In our model, infection probabilities in each time step depend on infection probabilities in the previous time step and one step transition probability matrix. In general, the transition probability matrix can be time variant or time invariant. If this matrix's elements are functions of infection probabilities in previous step, the corresponding dynamical system is non linear . In our model, we construct the one step transition matrix at each time step based on maximum likelihood estimation using observations.

We especially focus on building a dynamic network to predict the dispersal pattern of soybean rust. In this case, the nodes of the network are the counties of USA and we aim at predicting how much likely each county is to get infected over time based on observations of soybean rust up to the current time as well as previous years' corresponding observations. Soybean and kudzu densities in each county, daily wind data, and distance between counties are used to build the dynamic network.

2.1.2 Discrete Time Markov Chain

There are two different approaches to model epidemic in networks with discrete time Markov chain. The first one is to model it using a discrete time Markov chain with a state vector which is the nodes' states at each time. So, the state vector can be written as:

$$X(n) = [X_1(n)X_2(n).....X_N(n)], \tag{2.1}$$

where $X(n)$ is state vector of chain at time step n, and $X_i(n)$ denotes the state of node i at time step n. $X_i(n)$ can be zero or one. Zero indicate that node i at time step n is healthy and one indicates that node i at time step n is infected. This Markov chain has 2^N different states and can transient from each specific state to another one at every time. One step transition probability matrix of this Markov chain is a $2^N \times 2^N$ matrix.
Another approach is modeling each node alone by a separate discrete time Markov chain. In this case instead of having one Markov chain with a $1 \times N$ state vector, we have N distinct Markov chains. Each of these N Markov chain has a scalar state which can be zero or one at each step. So, for the Markov chain of node i, we can write

$$X_i(n) = \{0, 1\}. \tag{2.2}$$

The one step transition probability matrix of node i at time step n is

$$P_i(n) = \begin{bmatrix} S_i(n) & 1 - S_i(n) \\ \delta_i & 1 - \delta_i \end{bmatrix}, \tag{2.3}$$

where $S_i(n)$ is the conditional probability that given node i is healthy at time step n, stays healthy at next time step; δ_i is the conditional probability that given node i is infected at time step n, is cured at next time step. The probability vector of states of node i can be written as:

$$\pi_i(n) = \begin{bmatrix} \pi_{i0}(n) & \pi_{i1}(n) \end{bmatrix}, \tag{2.4}$$

where $\pi_{i0}(n)$ is the probability that node i is healthy at time step n; $\pi_{i1}(n)$ is the probability that node i is infected at time step n. The probability vector of discrete time Markov chain at time step $n+1$ can be written in terms of that vector in previous time step and one step transition probability matrix

$$\pi_i(n+1) = \pi_i(n) P_i(n). \tag{2.5}$$

Based on (3.4) we can derive probability of infection of node i at time step $n+1$ as:

$$\pi_{i1}(n+1) = 1 - \pi_{i0}(n) S_i(n) - \pi_{i1}(n)\delta_i. \tag{2.6}$$

In general, it is assumed that the curing process of each node is independent of other nodes in the network. δ_i as a conditional probability is independent of states of other nodes. This probability can changes over time regarding the conditions of node i. However, we assume in our network this probability is constant and not function of time. The situation of $S_i(n)$ is completely different from δ_i. $S_i(n)$ is the conditional probability that node i does not get infection given this node is healthy at time step n. The definition of this probability shows that it depends on the states of other nodes in the network. We can write $S_i(n)$ as

$$\begin{aligned} S_i(n) &= P(X_i(n+1) = 0 | X_i(n) = 0) \\ &= \sum_{x_{N_i(n)}} [P(X_{N_i(n)}(n) = x_{N_i(n)}(n) | X_i(n) = 0)(1 - \beta_{i, x_{N_i(n)}}(n))], \end{aligned} \tag{2.7}$$

where $N_i(n)$ denotes the set of neighbors of node i at time n; $X_{N_i(n)}(n)$ is the states of neighbors of node i at time step n; $\beta_{i,x_{N_i(n)}}(n)$ is the conditional probability that node i get infection at next time step given node i is healthy at time step n and its neighbors state is $x_{N_i(n)}(n)$. $N_i(n)$ can change over time based on the condition of network. In dynamic networks, the adjacency matrix of network changes over time. So, for every nodes in dynamic networks, $N_i(n)$ is a function of time. $P(X_{N_i(n)}(n) = x_{N_i(n)}(n)|X_i(n) = 0)$ as conditional joint probability characterizes the spatial dependence of node i to its neighbors due to network topology. This probability is very difficult to derive. There are different ways to simplify the determination this joint probability such as independent model or Markov model. First, we talk about independent model.

Independent Model

In independent model, we assume that states of all nodes at each time step are spatially independent. If node i has k neighbors, the total number of states needed to describe the joint probability $P(X_{N_i(n)}(n) = x_{N_i(n)}(n)|X_i(n) = 0)$ is reduced from $O(2^k)$ to $O(k)$. So, we can write

$$\begin{aligned} &P(X_{N_i(n)}(n) = x_{N_i(n)}(n)|X_i(n) = 0) = P(X_{N_i(n)}(n) = x_{N_i(n)}(n)) \\ &= \prod_{j\in N_i(n)} P(X_j(n) = x_j(n)). \end{aligned} \tag{2.8}$$

$\beta_{i,x_{N_i(n)}}(n)$ as the conditional probability that node i gets infection at next time step given node i is healthy at time step n and its neighbors state is $x_{N_i(n)}(n)$. Based on independent assumption, we can write

$$\beta_{i,x_{N_i(n)}}(n) = 1 - \prod_{j\in N_i(n)} (1 - \beta_{ji}(n))^{x_j(n)}, \tag{2.9}$$

where $\beta_{ji}(n)$ is the probability that node i gets infected at time step $n+1$ if node j is infected at time step n. Using (3.9) and (3.10), we can write $S_i^{ind}(n)$ as

$$S_i^{ind}(n) = \prod_{j\in N_i(n)} (1 - \beta_{ji}(n)P(X_j(n) = 1)) = \prod_{j\in N_i(n)} (1 - \beta_{ji}(n)\pi_{j1}(n)), \tag{2.10}$$

where $\pi_{j1}(n)$ is the probability that node j is infected at time n. Node i one step transition probability matrix considering independent model can be written as

$$\begin{aligned} P_i^{ind}(n) &= \begin{bmatrix} S_i^{ind}(n) & 1 - S_i^{ind}(n) \\ \delta_i & 1 - \delta_i \end{bmatrix} \\ &= \begin{bmatrix} \prod_{j\in N_i(n)} (1 - \beta_{ji}(n)\pi_{j1}(n)) & 1 - \prod_{j\in N_i(n)} (1 - \beta_{ji}(n)\pi_{j1}(n)) \\ \delta_i & 1 - \delta_i \end{bmatrix}. \end{aligned} \tag{2.11}$$

As you can figure out from (2.11), $P_i^{ind}(n)$ does not depend on $\pi_i(n)$. So, the Markov chain of node i is a linear dynamical system whose state vector is $\pi_i(n)$. The state equation of this linear dynamical system can be written as

$$\pi_i(n+1) = \pi_i(n)P_i^{ind}(n). \tag{2.12}$$

Constraint on $\beta_{ji}(n)$

$\beta_{ji}(n)$ is the probability that node i gets infected at time step $n+1$ if node j is infected at time step n. So, $\beta_{ji}(n)$ as a probability must be between zero and one. We can get the same result from (2.10). $S_i^{ind}(n)$ as an element of one step transition probability is between zero and one. So, We can write

$$0 \leq S_i^{ind}(n) \leq 1 \Rightarrow 0 \leq \prod_{j \in N_i(n)} (1 - \beta_{ji}(n)\pi_{j1}(n)) \leq 1 \tag{2.13}$$

$$\Rightarrow 0 \leq (1 - \beta_{ji}(n)\pi_{j1}(n)) \leq 1 \Rightarrow 0 \leq \beta_{ji}(n)\pi_{j1}(n) \leq 1. \tag{2.14}$$

Because $0 \leq \pi_{j1}(n) \leq 1$, we can conclude that $0 \leq \beta_{ji}(n) \leq 1$.

2.1.3 Modeling Soybean Rust

We can consider different time steps for our network model like daily, weekly, biweekly, or monthly. Because there is a lag of 7 to 14 days since spores reach one county and the time they make infection in that county, we have chosen biweekly time steps. There are different factors which affect the dispersal of soybean rust every year. These factors include the direction and speed of wind from each county toward the other counties, the distance between counties, and soybean and kudzu area in each county. We consider the effects of these factors in our model by determining $\beta_{ji}(n)$ based on them. We can write $\beta ji(n)$ as

$$\beta_{ji}(n) = \frac{\theta_{ij} d_i d_j w_{ji}(n)}{L_{ji}^2}, \tag{2.15}$$

where $d_i(km^2)$ is the total area of soybean and kudzu in county i; $d_j(km^2)$ is the total area of soybean and kudzu in county j; $w_{ji}(n)$ (kilometer per hour) is the maximum wind speed projection from county j toward county i during the last two weeks; $L_{ij}(km)$ is distance between counties i and j; θ_{ji} is the parameter which we need to estimate using observations of soybean rust and can be different for each link from county i toward county j. d_i, d_j, and L_{ij} are constant. However, $w_{ji}(n)$ changes over time and causes $\beta_{ji}(n)$ to be a function of time. There is this possibility that wind at county j do not have any projection toward county i or the projected wind speed is negative. In these cases, we consider w_{ji} zero for those days. Being functions of time of w_{ji}s makes the network be dynamic. Therefore, the network has a weighted adjacency matrix that changes with time. Substituting $S_i^{ind}(n)$ in (3.5), we can write

$$\begin{aligned}\pi_{i1}(n+1) &= 1-\pi_{i0}(n)S_i^{ind}(n)-\pi_{i1}(n)\delta_i \\ &= 1-\pi_{i0}(n)\prod_{j\in N_i(n)}(1-\beta_{ji}(n)\pi_{j1}(n))-\pi_{i1}(n)\delta_i. \end{aligned} \tag{2.16}$$

Pluging $\beta_{ji}(n)$ into (2.16), $\pi_{i1}(n+1)$ can be written as

$$\pi_{i1}(n+1) = 1-\pi_{i0}(n)\prod_{j\in N_i(n)}(1-\frac{\theta_{ji}d_i d_j w_{ji}(n)}{L_{ji}^2}\pi_{j1}(n))-\pi_{i1}(n)\delta_i, \tag{2.17}$$

The likelihood function of state of node i at time $n+1$ given θ_{ji}s can be written as

$$\begin{aligned}P(X_i(n+1)) &= X_i(n+1)\pi_{i1}(n+1)+(1-X_i(n+1))(1-\pi_{i1}(n+1) \\ &= X_i(n+1)+(1-2X_i(n+1))(1-\pi_{i1}(n+1)). \end{aligned} \tag{2.18}$$

Estimation of θ_{ji}

To build the network, we need to estimate θ in order to calculate π_{i1} at each time step. There are two categories of estimation method: Bayesian and non Bayesian estimation. For Bayesian estimation, the distributions of parameters which are estimated are required. In our problem, distribution of θ_{ji} is unknown. So, we employ non Bayesian methods to estimate them. The two common non Bayesian estimators are maximum likelihood estimator (MLE) and minimum variance unbiased estimator (MVUE). Nyman Fisher factorization theorem shows that we cannot extract sufficient statistic from likelihood function(2.18). Therefore, MVUE is not applicable to this network and we use MLE to estimate θ_{ji}.

MLE

According to MLE approach, the estimate of θ_{ji} is the value that maximizes the likelihood function given observations. So, we can write

$$\widehat{\theta}_{ji,MLE} = argmax_{\theta_{ji}}\left\{P(X_i(n+1))\right\}. \tag{2.19}$$

The method of finding $\widehat{\theta}_{ji,MLE}$ depends on properties of $P(X_i(n+1))$. In general, there are three types of likelihood functions: (1) likelihood is a linear function of parameters, (2) likelihood is a convex or concave function of parameters, and (3) likelihood is nonlinear function of parameters and neither convex nor concave function of parameters. On the other hand, $P(X_i(n+1))$ changes over time and depends on some time-varying factors such as $w_{ji}(n)$, $\pi_{i1}(n)$, and $\pi_{j1}(n)$. It causes likelihood function at each time step can belong to one of three different mentioned categories. When the observation for node i at time step $n+1$, $X_i(n+1)$, is zero, $P(0)$ can be written

$$P(0) = \pi_{i0}(n) \prod_{j \in N_i(n)} (1 - \frac{\theta_{ji} d_i d_j w_{ji}(n)}{L_{ji}^2} \pi_{j1}(n)) + \pi_{i1}(n)\delta_i, \tag{2.20}$$

(2.20) shows that the maximum likelihood solution is zero when the corresponding observation is zero as well. Therefore, $\widehat{\theta}_{ji,MLE} = 0$ because replacing θ_{ji} by zero, maximizes $P(0)$. In the other hand, When the observation for node i at time step $n+1$, $X_i(n+1)$, is one, $P(1)$ can be written

$$P(1) = 1 - \pi_{i0}(n) \prod_{j \in N_i(n)} (1 - \frac{\theta_{ji} d_i d_j w_{ji}(n)}{L_{ji}^2} \pi_{j1}(n)) - \pi_{i1}(n)\delta_i, \tag{2.21}$$

(2.21) shows that $\widehat{\theta}_{ji,MLE} = \frac{L_{ji}^2}{d_i d_j w_{ji}(n)}$ maximizes $P(1)$ when $\pi_{j1}(n)$ is not zero. Moreover, if $\widehat{\theta}_{ji,MLE} = 0$ maximizes $P(1)$ when $\pi_{j1}(n) = 0$. It is remarkable that $P(X_i(n+1))$ as the likelihood function changes over time and is district for each node. It is reasonable to expect that these maximum likelihood estimators change with time by variation in $w_{ji}(n)$, $\pi_{i1}(n)$, and $\pi_{j1}(n)$.

2.2 Estimation of Missing Observations

2.2.1 Introduction

We aim at estimating θ_{ji} at each time step for every node so that observations at each time step for every node are required as well. However, observations of soybean rust for every county at every day are not available. Soybean rust observations exist for only a few number of counties and a few days. Therefore, estimation of missing is inevitable.

2.2.2 Observation Assumptions

We consider two important assumptions to use available observations to estimate missing observations. First, if there is a healthy observation for one county at a specific day, we assume that county has been healthy up to that day. However, we cannot assume that the county is healthy also after that day by only taking into account the observation at that day. Second, if there is a infection observation for one county at a specific day, we assume that county remains infected after that observation. However, we cannot determine the status of that county before that day. There are four different types of county based on the available observations.

Type 1 : for this type of counties, there is not any observation at all. So, we use other counties' available observations to estimate their missing observations.

Type 2 : There are only healthy observations for this type of counties. First, we find the last available observation for each county. Then, we assume that the county has been healthy for all days before the last observation. Finally, we estimate the state of the county

for all days after the last observation.
Type 3 : There are just infection observations for this type of counties. First, we find the first infection observation for each county. Second, we assume that the county is infected for all days after the first infection observation. We estimate the state of the county for all days before the first observation.
Type 4 : For this type of counties, there are both of healthy and infection observations. It is assumed that an infected county cannot get cured. Therefore, we expect the last healthy observation is before the first infection observation for each of these counties. First, we find the last healthy observation and the first infection observation for each county. Then, we assume that the county is healthy for all days before its last healthy observation and infected for all days after its first infection observation. Finally, we estimate the state of the county for the days between the last healthy observation and the first infection observation. Among 2005's reported observations, for two counties the last day of healthy observation is after the first day of infection observation, which is different from our assumption that a county cannot get cured when it gets infected. It is because that these counties have both of soybean and kudzu and in some days there are healthy observations for one of these plants and at the same time infection observations for another one. Because in our model we do not distinguish between soybean and kudzu, we assume counties are infected if one of these two plants gets infected. In this case, we consider healthy observations for all days before the first infection observation, and infection observations for all days after the first infection observation.

As we mentioned earlier, daily observations are required to estimate θ. In first step, we just employ the assumption that a county cannot be cured when it gets infected to determine the daily observations of each county without estimation of missing observations. As result of the first step, there are some available daily observations and some missing observations for each county over time. For example, we plot the daily observations for 20^{th} of each month of 2005 in Fig. 2.1 to Fig. 2.12 without estimation of missing observations. Black dots represent the counties with missing observations. Green dots represent the counties with healthy daily observations. Red dots represent the counties with infection daily observations. As you can see in these figures, there are so many missing observations for the counties over time.

The number of infection observations start to increase since July. In the last two months of 2005, all of the observations are infection and there is not any healthy observation.

In the second step, we estimate the missing daily observations. There are different methods to estimate the missing observations and they can be divided into two main groups: (1) those methods which use just available daily observations at each time step to estimate the missing observations and they do not use available observations of pervious days. We name this type of methods the spatial methods. (2) This type of methods unlike the first group use the available daily observation of previous days as well as the current day's available data to estimate the current day's missing daily observations. We name this groups the spatial temporal methods.

The first type of methods does not show good performance in our problem. It is because that if we limit ourselves just to use the current day available observations to estimate the missing

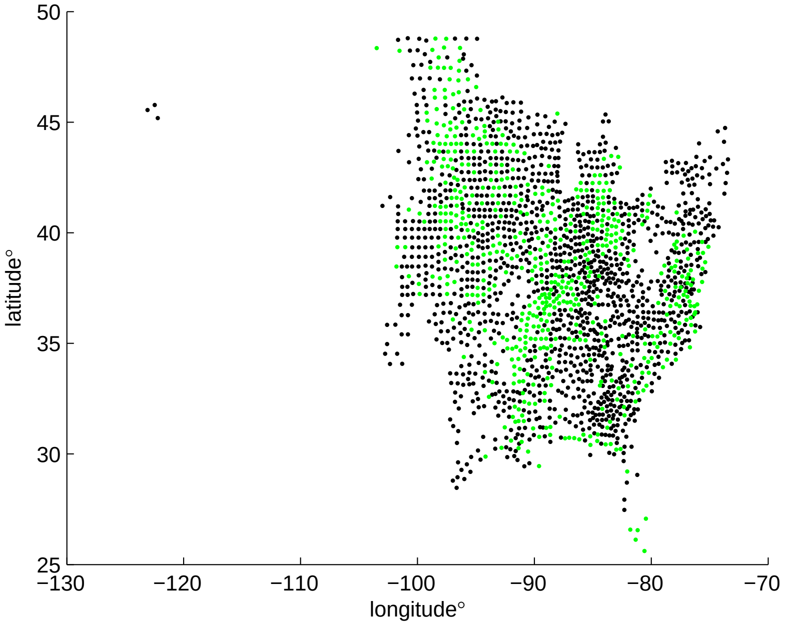

Figure 2.1: *Available observations on January* 20^{th}.

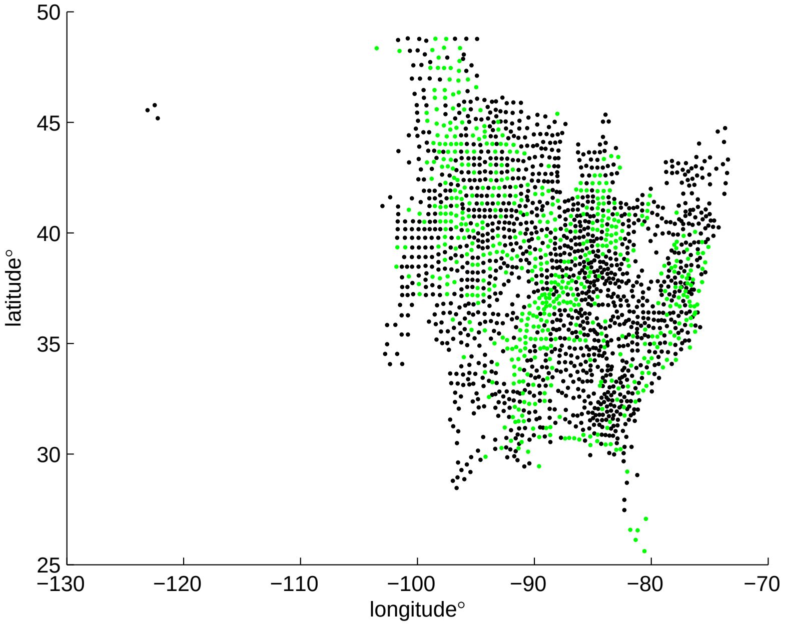

Figure 2.2: *Available observations on February* 20^{th}.

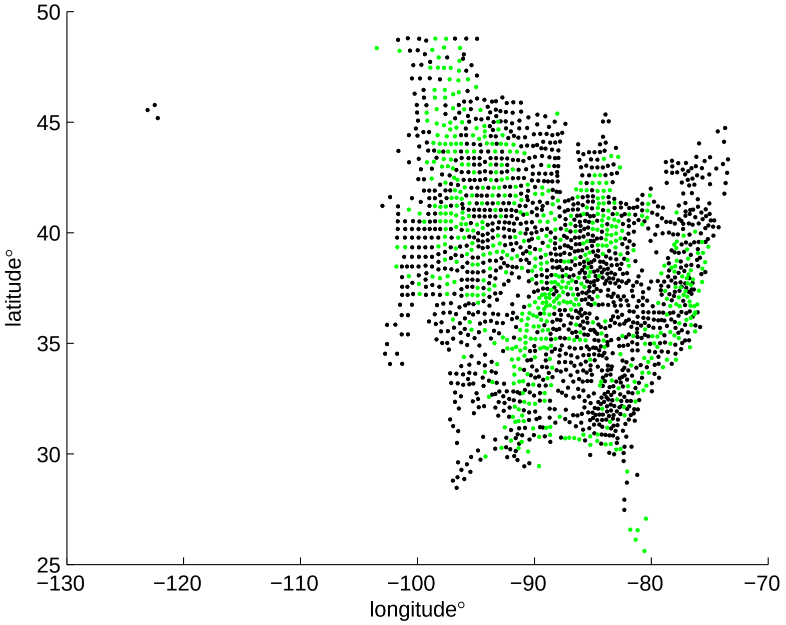

Figure 2.3: *Available observations on March* 20^{th}.

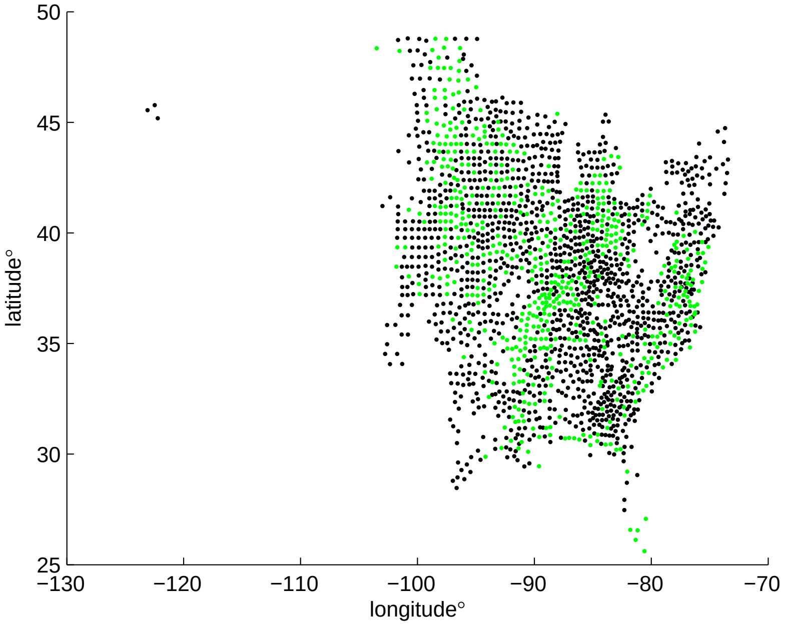

Figure 2.4: *Available observations on April* 20^{th}.

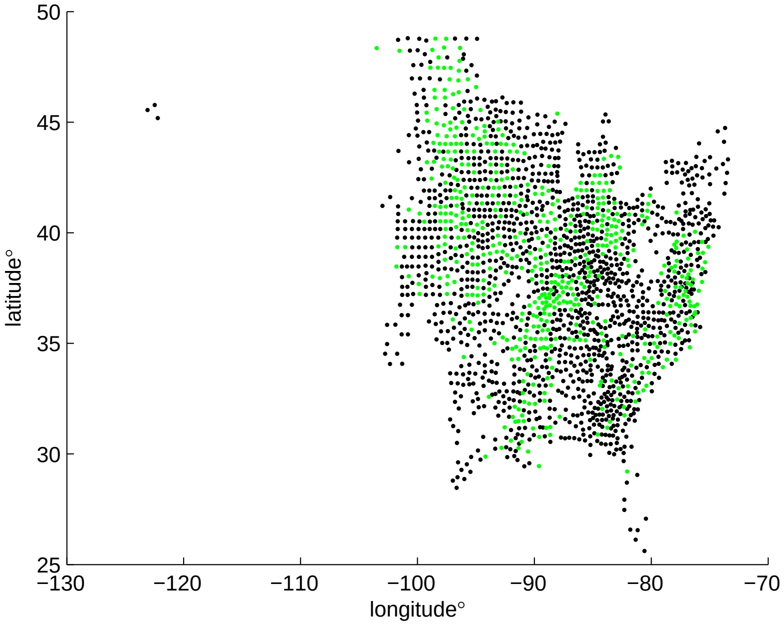

Figure 2.5: *Available observations on May* 20^{th}.

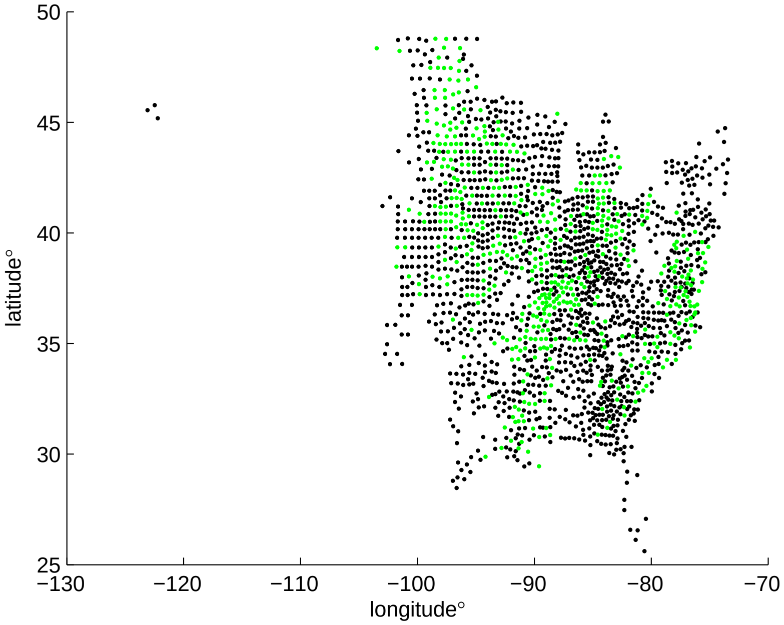

Figure 2.6: *Available observations on June* 20^{th}.

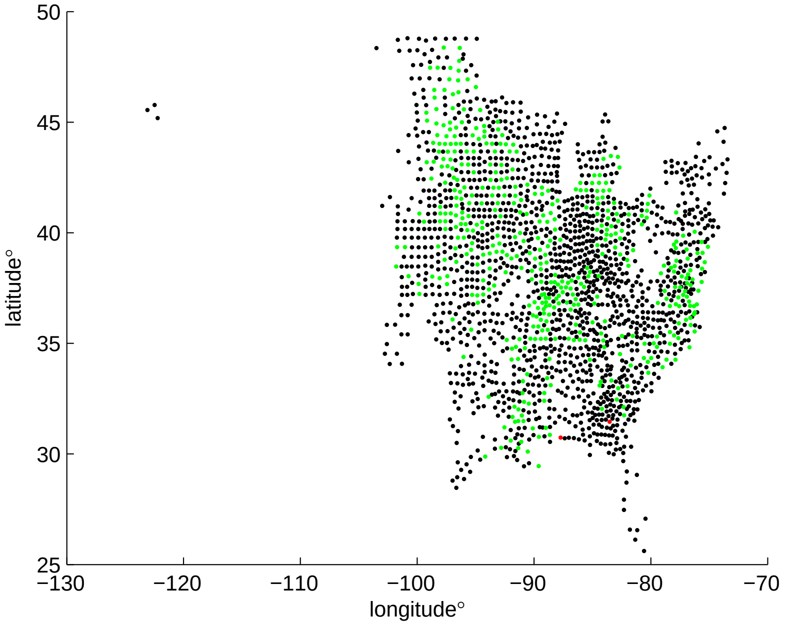

Figure 2.7: *Available observations on July* 20^{th}.

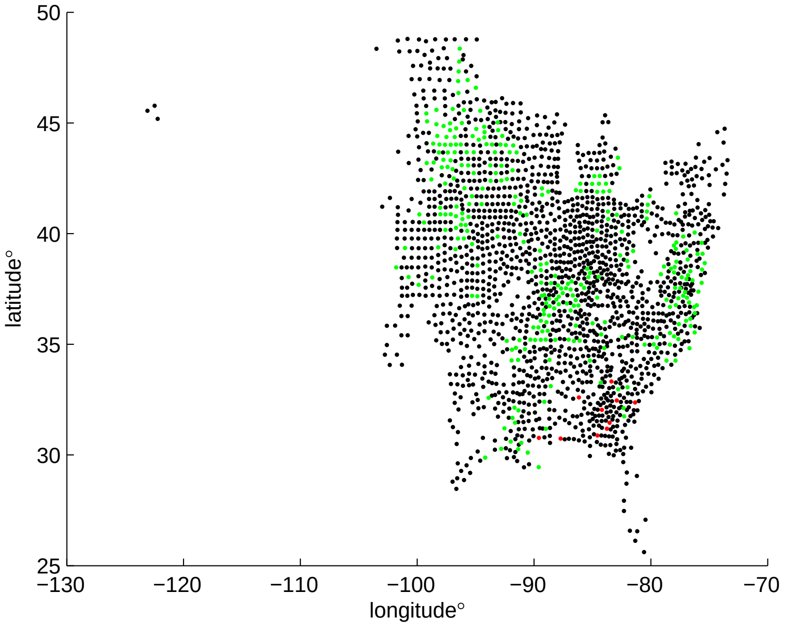

Figure 2.8: *Available observations on August* 20^{th}.

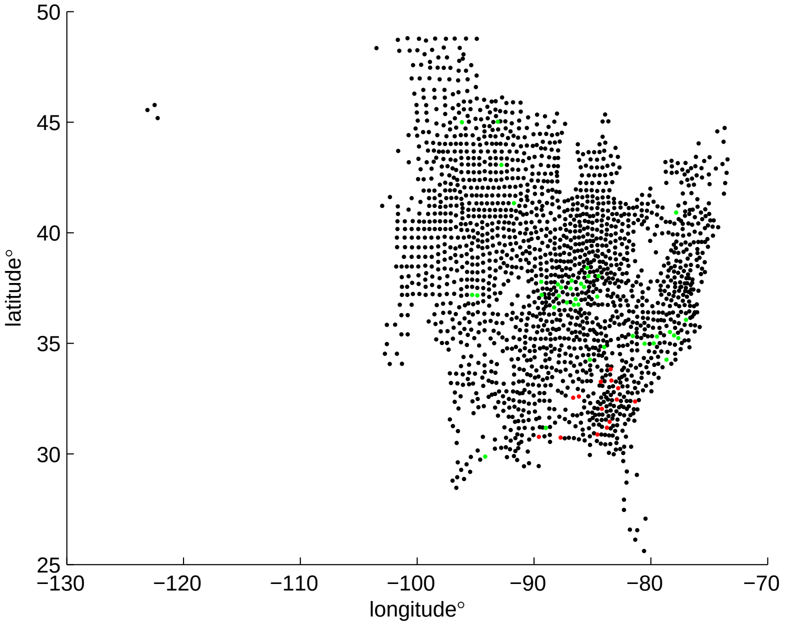

Figure 2.9: *Available observations on September* 20^{th}.

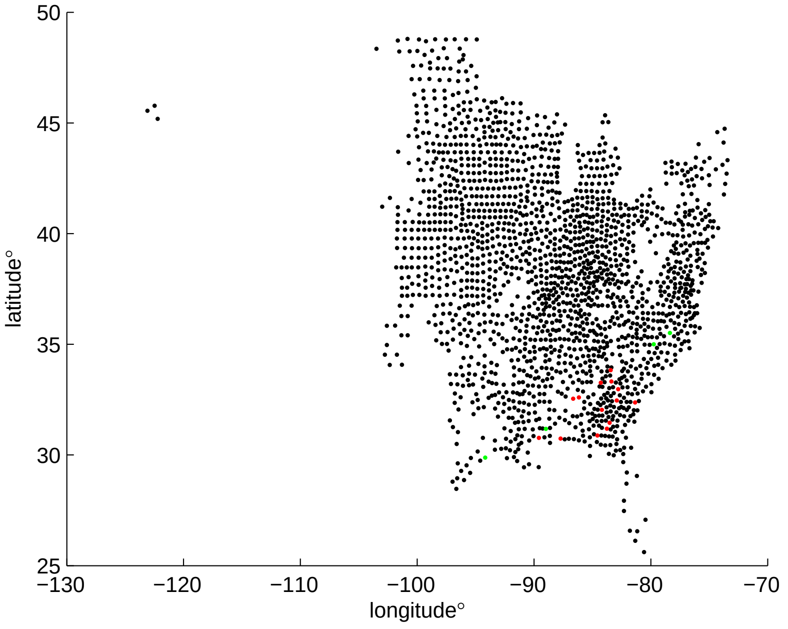

Figure 2.10: *Available observations on October* 20^{th}.

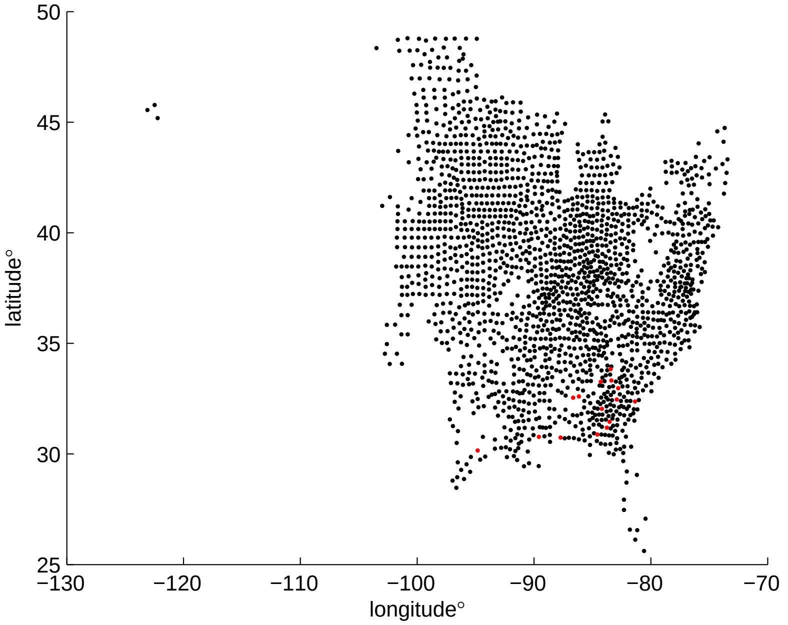

Figure 2.11: *Available observations on November* 20^{th}.

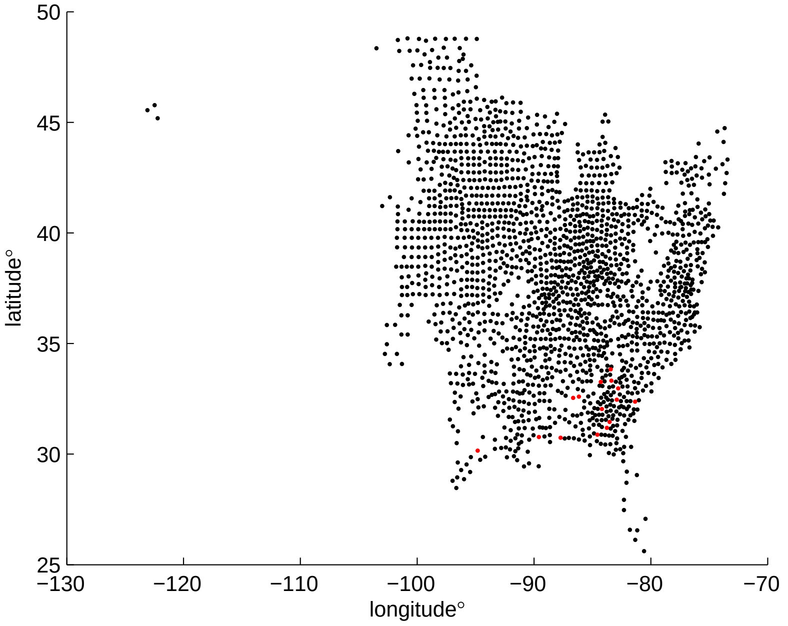

Figure 2.12: *Available observations on December* 20^{th}.

observations, there are some days especially after September that there is not any available daily observation in reasonable distance of the majority of counties. For example, we are sure that counties in the north of US have been healthy throughout the year. However, in the last three months of year there is not any available observation at north counties. If we use the current day's daily available observations, it means that we have to use south counties' observation that all of them are infection observations to estimate north counties' missing daily observations. This results that estimated observations for all north counties in the last three months become infected.

2.2.3 Spatial Temporal estimator of missing observations with equal gains

In this method first for estimation of missing observation of county i at time step n, we find the counties whose distances to that county are less than $200km$. to simplify the calculations, we assign -1 to available healthy daily observations unlike before which we denote healthy observations by 0 (We change healthy observation again to 0 later to use them in our probabilistic model). We use 1 to indicate infection observations same as before. Finally, we denote missing observations by 0. The estimated value of missing observation of county i at time step n can be written as

$$\widehat{Y}_i(n) = \begin{cases} 1 & \left(\sum_{k=1}^{n} \sum_{j \in M_i} Y_j(k)\right) \geq 0 \\ -1 & \left(\sum_{k=1}^{n} \sum_{j \in M_i} Y_j(k)\right) < 0 \end{cases}, \tag{2.22}$$

where M_i denotes the counties whose distances to county i are less than $200km$; $Y_j(k)$ is the daily observation of node j at time step k which is -1 if it is available and healthy, 1 if it is available and infection, and 0 if it is not available. We do not use estimated values of missing observations to estimate other missing observations at next time steps or the current time step. (2.22) shows that this method gives the equal gains to all of observation regardless of their dates. It means that $Y_j(n)$ and $Y_j(n-1)$ play the same role in determining $\widehat{Y}_i(n)$.
We plot the estimated values calculated by (2.22) for missing observations of 20^{the} September of 2005 in Fig. 2.13. Comparing Fig. 2.13 with Fig. 2.9, you can see that the estimated values for all missing observation(black dots in Fig. 2.9) are healthy is represented by green dots in Fig. 2.13. We expect that estimated values for south east counties' missing observations become +1 (infection) because all the available observation for that part of the US in 20^{th} September are infection, which is different from the output of the estimator. It is because that this estimator gives equal gains to all observation regardless of their dates, which is not reasonable. To solve this issue, we propose an estimator which give more weights to newer observations in the next section.

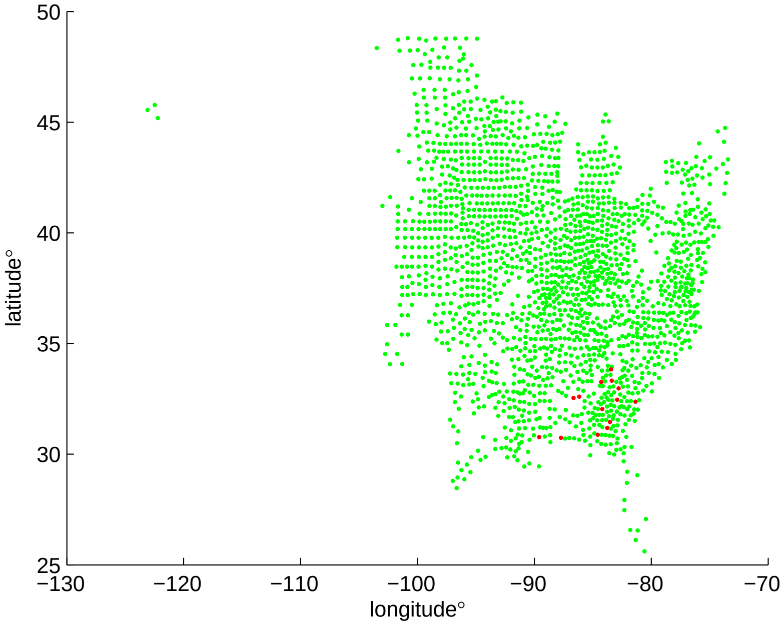

Figure 2.13: *Estimated Missing observations using Equal Gains on September* 20^{th}.

2.2.4 Spatial Temporal Estimater of missing observations with different gains

In this section we derive a new estimator in which the gain of each available observation in estimation depends on its date unlike the estimator with equal gains in section 2.2.3. By this way, the output of the estimator for one specific day mainly depends on that day's available observations compare to available observation of previous days. To estimate the missing observation of county i first, we find the counties whose distances to county i are less than $200km$. Then, we use their available observations for estimation of the missing observations of county i. We assign -1 to available healthy daily observations. We denote infection and missing observations, respectively, by 1 and 0. The function to estimate the value of missing observation of county i at time step n can be written as

$$\widehat{Y}_i(n) = \begin{cases} 1 & \left(\frac{\alpha}{n-1}\sum_{k=1}^{n-1}\sum_{j\in M_i} Y_j(k) + \alpha\sum_{j\in M_i} Y_j(n)\right) \geq 0 \\ -1 & \left(\frac{\alpha}{n-1}\sum_{k=1}^{n-1}\sum_{j\in M_i} Y_j(k) + \alpha\sum_{j\in M_i} Y_j(n)\right) < 0 \end{cases}, \tag{2.23}$$

where M_i denotes the counties whose distances to county i are less than $200km$; α is the gain factor which determines the importance of the current day's available observations; $Y_j(k)$ is the daily observation of node j at time step k which can be -1 if it is available and healthy, 1 if it is available and infection, and 0 if it is not available. We do not use estimated values of missing observations at time step n for estimate of missing observations at next time steps. We examined different values of α such as .5, .75, and .9 and found out that

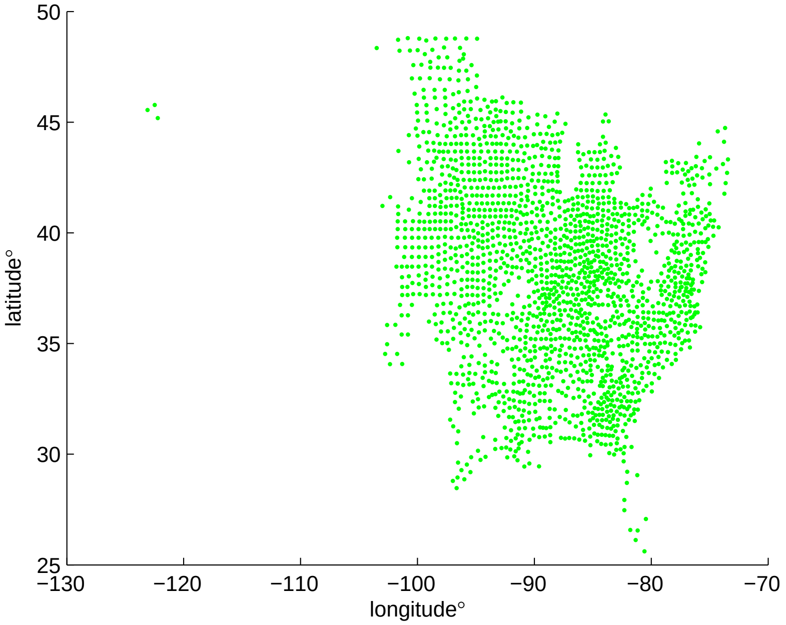

Figure 2.14: *Estimated missing observations using α gain on January 20^{th}.*

we can get the best results with $\alpha = .9$. We plot the estimated missing observations using (2.23) on 20^{th} of each month of 2005 in Fig. 2.14 to Fig. 2.25. Comparing Fig. 2.22 with Fig. 2.9, it can be seen that the most estimated values for missing observations of south east counties are infection, which is reasonable.

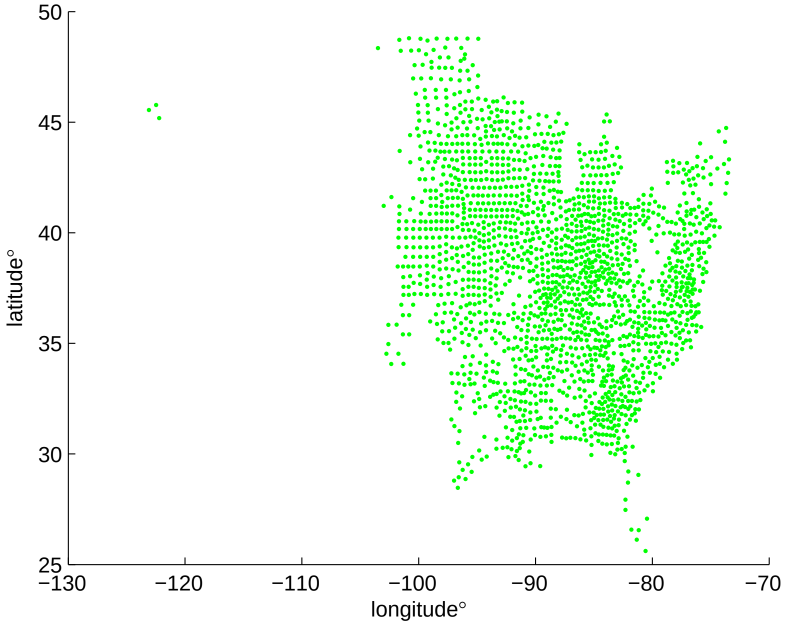

Figure 2.15: *Estimated missing observations using α gain on February* 20^{th}.

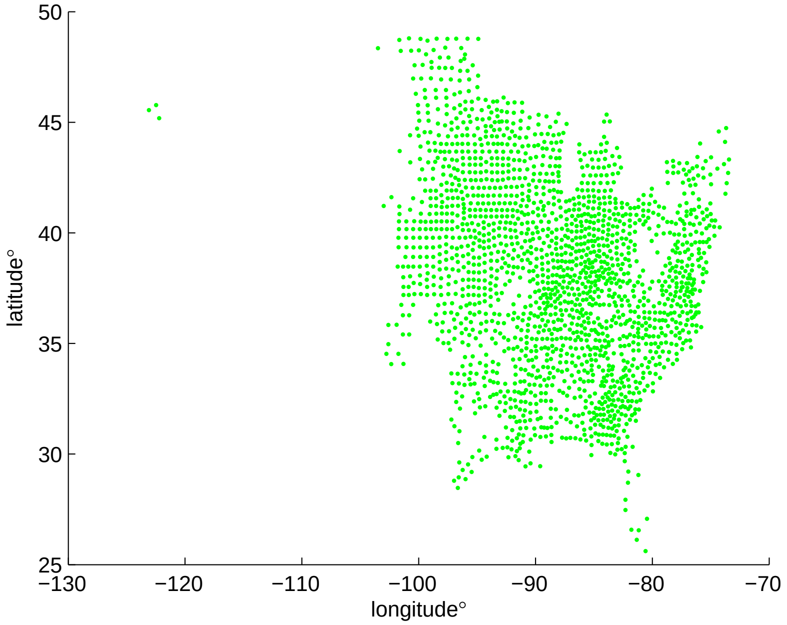

Figure 2.16: *Estimated missing observations using α gain on March* 20^{th}.

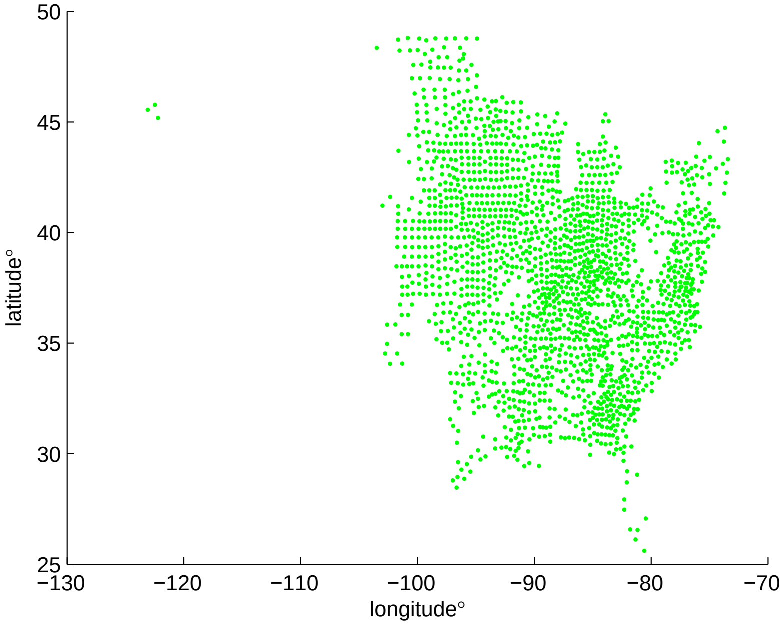

Figure 2.17: *Estimated missing observations using α gain on April 20th.*

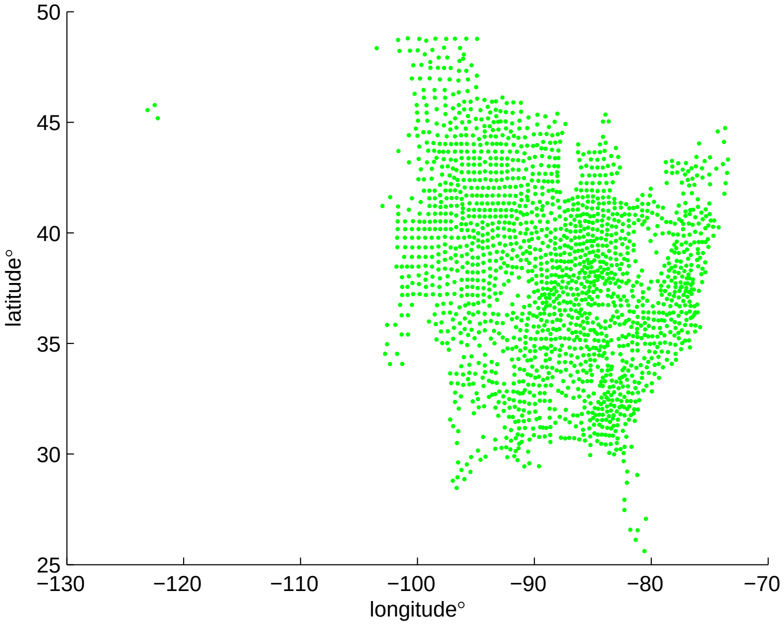

Figure 2.18: *Estimated missing observations using α gain on May 20th.*

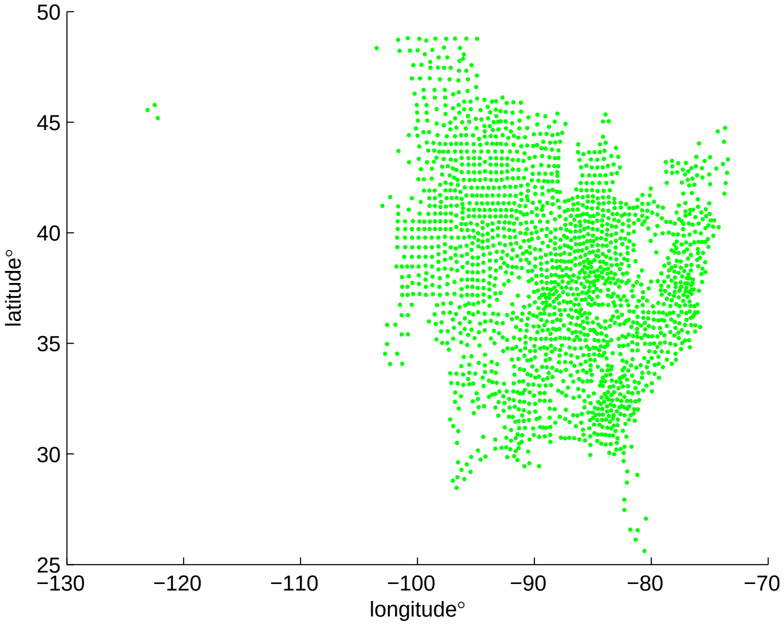

Figure 2.19: *Estimated missing observations using α gain on June 20^{th}.*

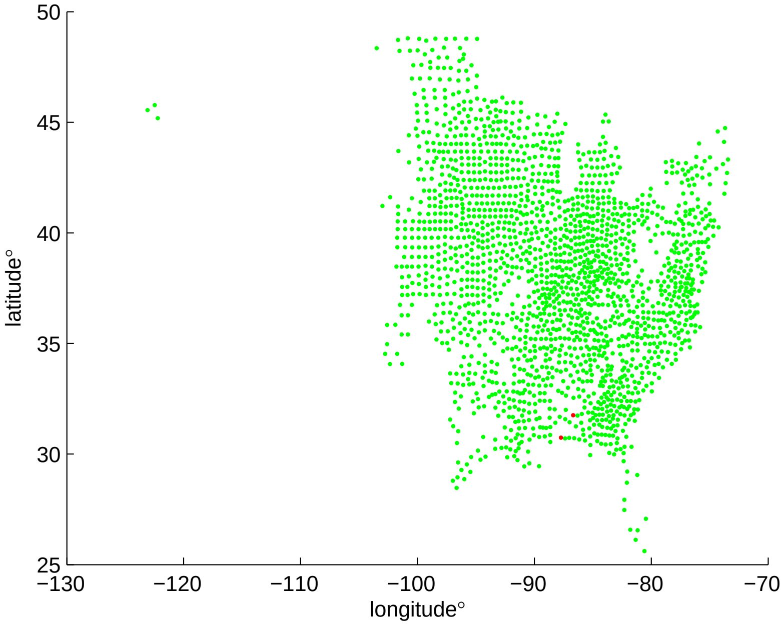

Figure 2.20: *Estimated missing observations using α gain on July 20^{th}.*

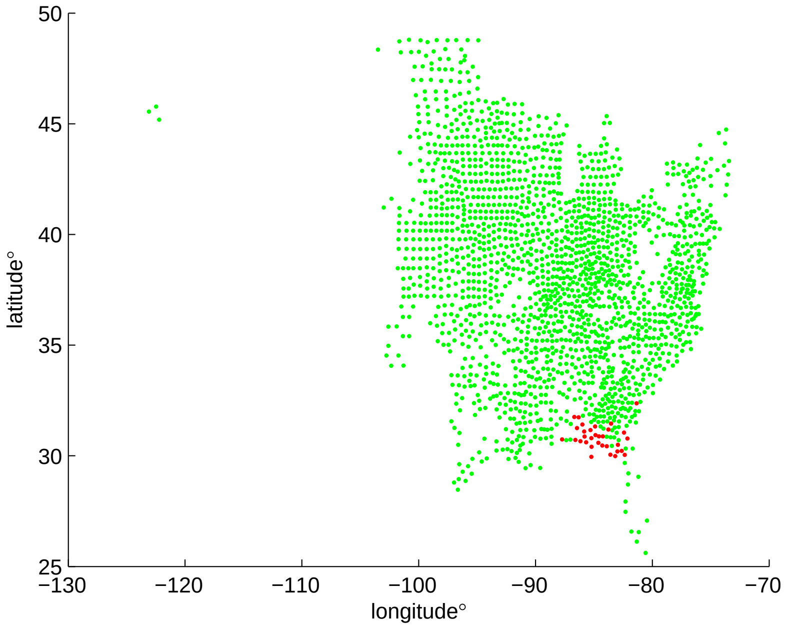

Figure 2.21: *Estimated missing observations using α gain on August 20^{th}.*

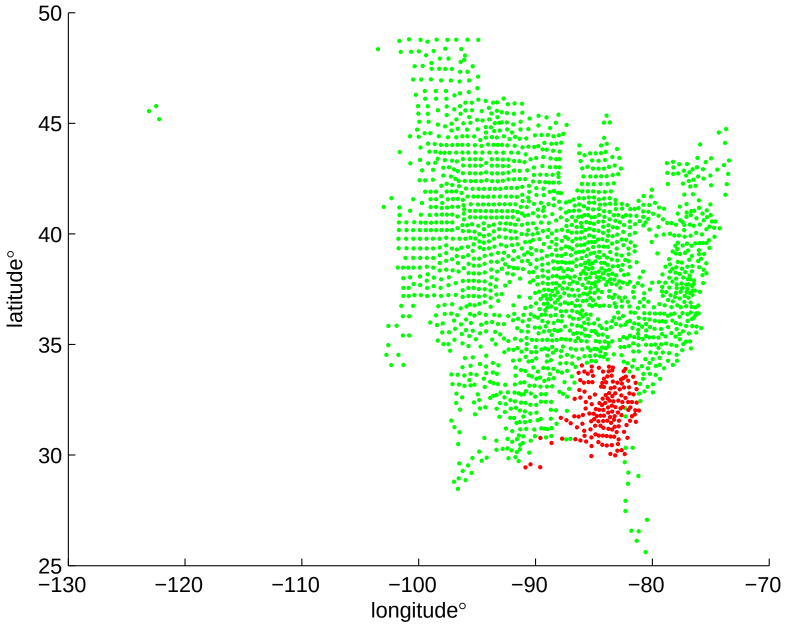

Figure 2.22: *Estimated missing observations using α gain on September 20^{th}.*

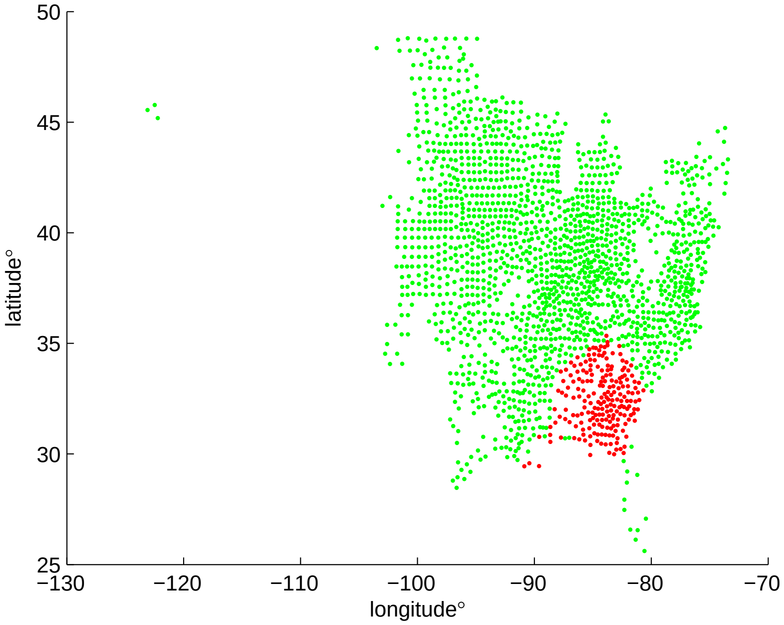

Figure 2.23: *Estimated missing observations using α gain on October* 20^{th}.

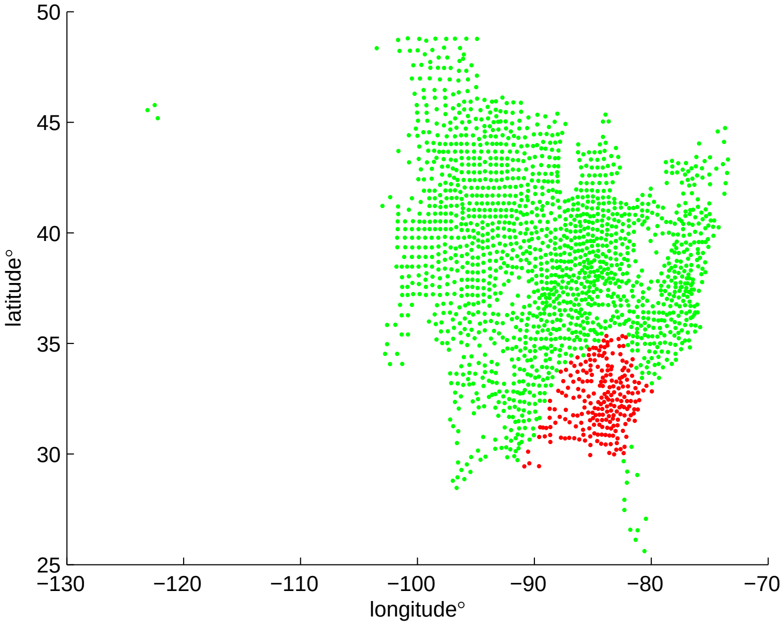

Figure 2.24: *Estimated missing observations using α gain on November* 20^{th}.

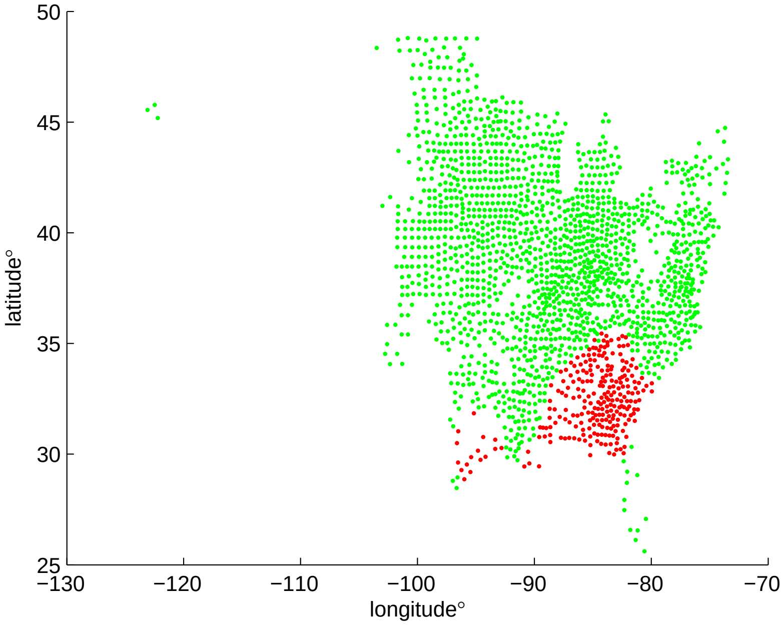

Figure 2.25: *Estimated missing observations using α gain on December 20^{th}.*

2.3 Simulation Results

In this section, first we compare our epidemic model predictions for soybean rust dispersal with its corresponding observations. Secondly, we state three different hypothesis regarding how to use years' data to estimate model's parameters and assess them based on simulation results. As we mentioned earlier, to predict one specific year's soybean rust dispersal, the rest of years' observations are used to build the corresponding network or in other word estimation of the network's parameters θ_{ji}s. For instance, assume that we want to predict disease spreading for year of 2007. In this case, we have different choices for how to build its network. One way to do it is only exploiting 2005's observations and data to estimate the network's parameters. If we estimate the network's parameters only using 2005's data, we don't exploit other years' data which can make our model more general. If we assume that our model is a causal system, we have to limit ourselves to use all the years' data before 2007. In this case, we have three different options: (1) only 2005's data (2) only 2006's data (3) data of 2005 and 2006 together because soybean rust data is available after 2005. Using 2005's data and observations, we obtain one version of the estimated network's parameters and if we use 2006's data and observations, we obtain different version of estimated network's parameters. So for the case of combination of 2005 and 2006 to build the network, one questions arises: how to combine these two different versions of estimated parameters? We average those two versions to obtain one set of parameters. If we leave the causality of the system, there are 15 permutations of 2005, 2006, 2008, and 2009 to build the network. It is noticeable that in the case of using more than two years to build the network, we average all over those years to come up with one set of parameters.
To predict disease dispersal for one specific year at a given date, we use also the observations for that specific year until that date, which increase the accuracy of the predictions. The other issue that we need to notice is that when we run our model to predict the dispersal of disease, it is possible that in some cases $\theta_{ij}\frac{d_i d_j w_{ji}(n)}{L_{ji}^2}$ become greater than one. For those cases, we consider one for the link weight. This can happen when a link's wind projection at one time step is greater than the value of this variable for the exact time step in the year which are used to build the model.
For example in 2005, Soybean rust started after the mid of May. Fig. 2.26 shows the Soybean rust observation in the second portion of May 2005 and Fig. 2.27 plots the model outputs for that interval time. Black dots represent the healthy counties and red dots represent infected counties. In figures from 2.28 up to 2.41, you can see the Soybean rust observations for the second part of April until December and the corresponding predictions. As it can be seen in these figures, the network's predictions match more with the observations for the last months of 2005 compare to the first months of infection. It is because of that over time we feed the model more data by providing it with the current year's observations.

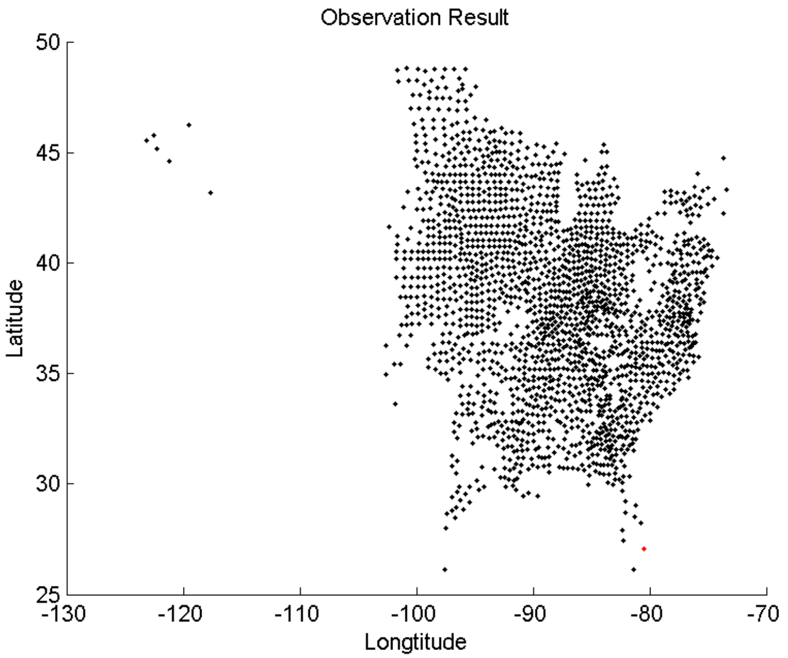

Figure 2.26: *Observations in the last two weeks of May.*

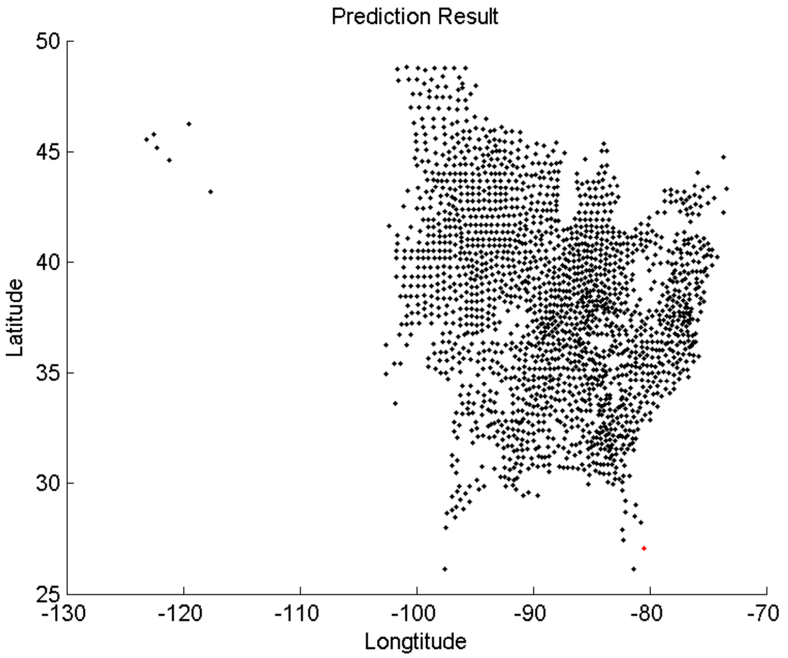

Figure 2.27: *Prediction for the last two weeks of May.*

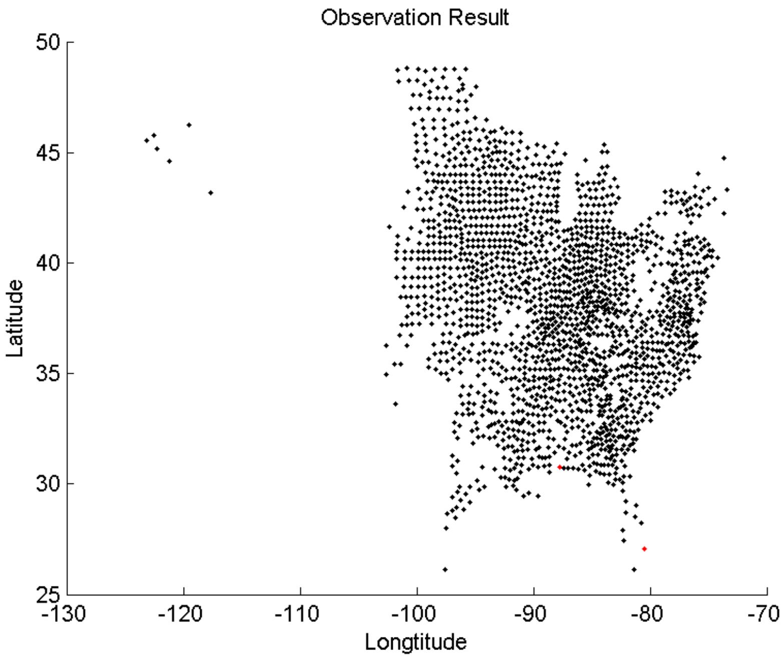

Figure 2.28: *Observations in the last two weeks of June.*

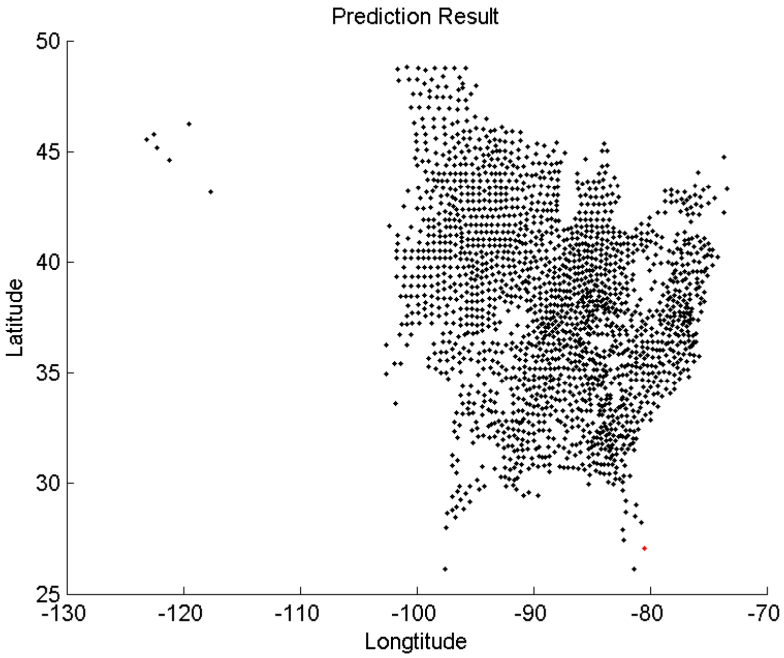

Figure 2.29: *Prediction for the last two weeks of June.*

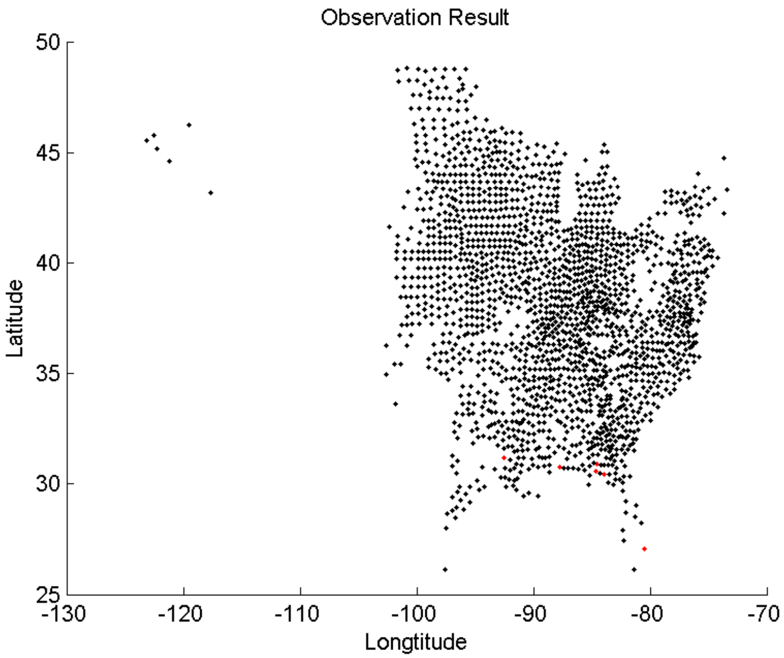

Figure 2.30: *Observations in the last two weeks of July.*

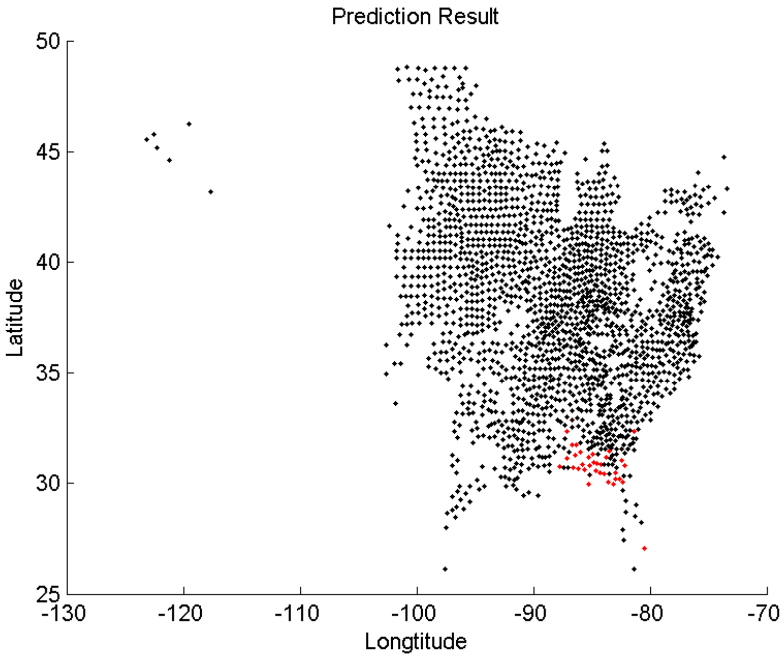

Figure 2.31: *Prediction for the last two weeks of July.*

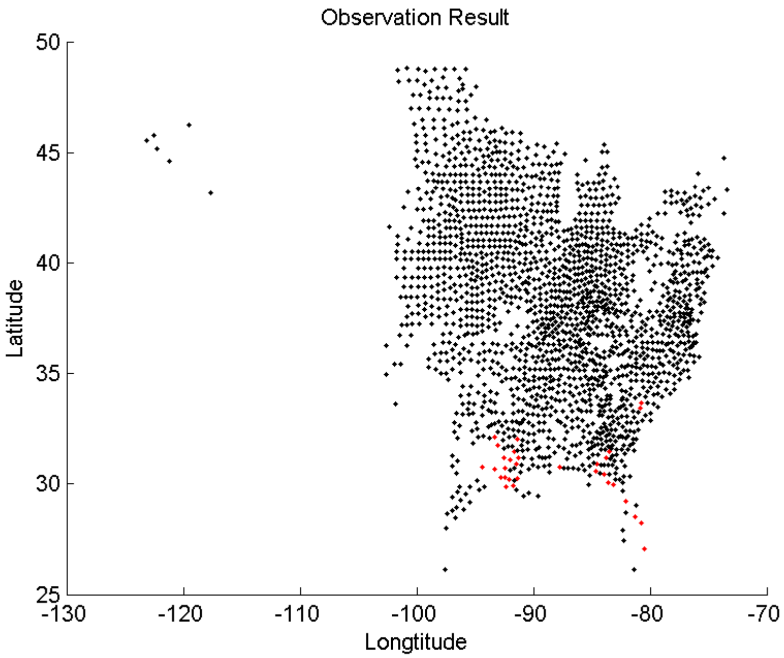

Figure 2.32: *Observations in the last two weeks of August.*

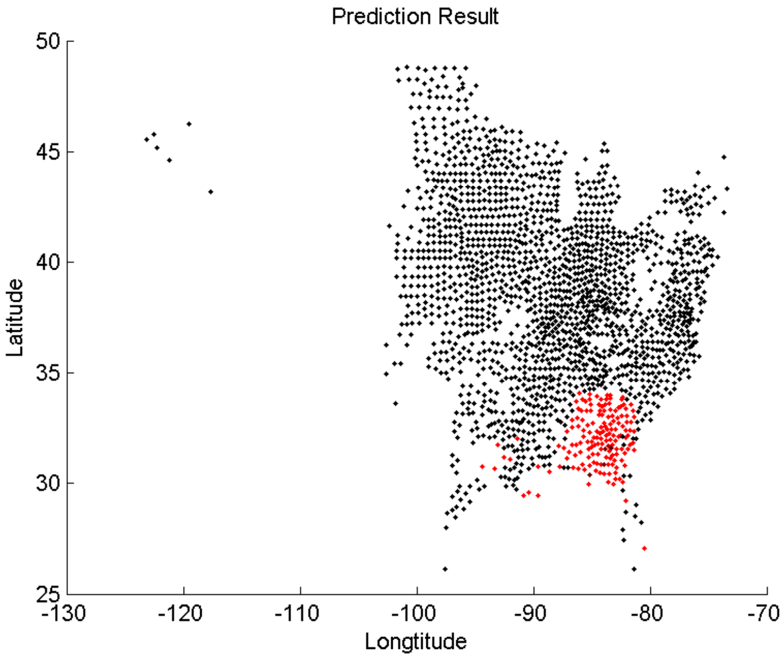

Figure 2.33: *Prediction for the last two weeks of August.*

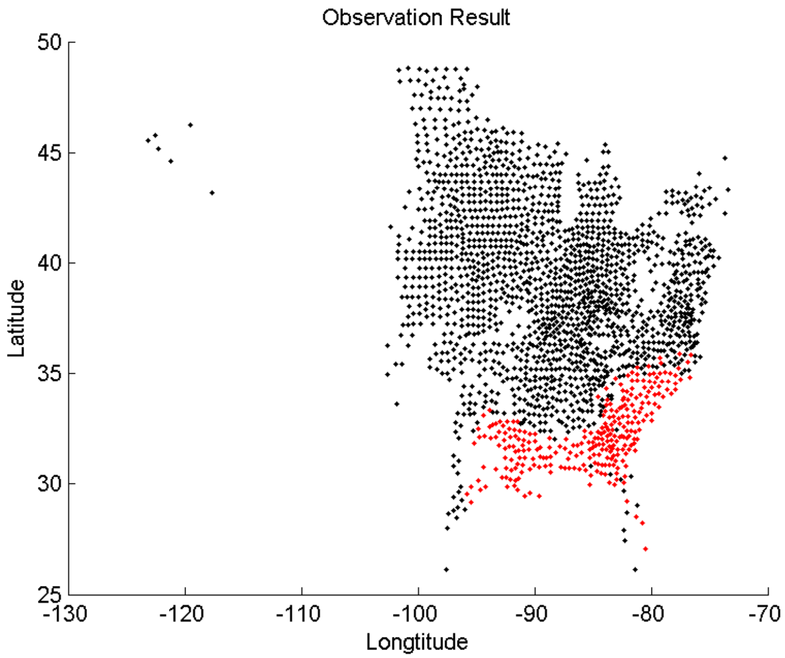

Figure 2.34: *Observations in the last two weeks of September.*

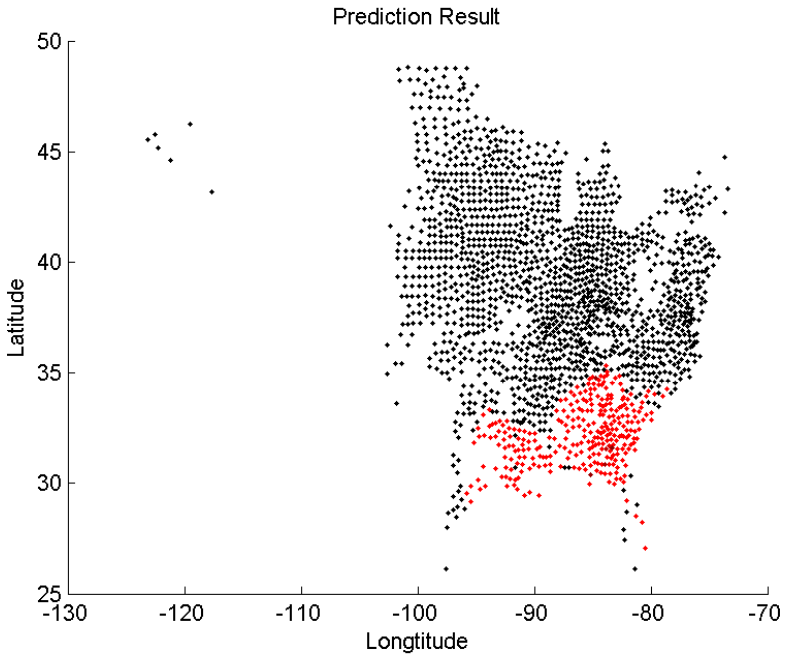

Figure 2.35: *Prediction for the last two weeks of September.*

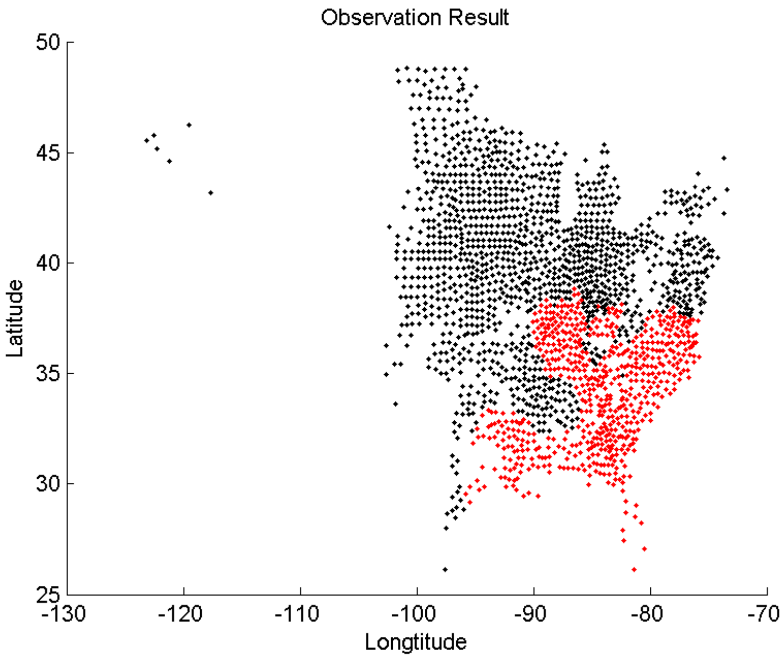

Figure 2.36: *Observations in the last two weeks of October.*

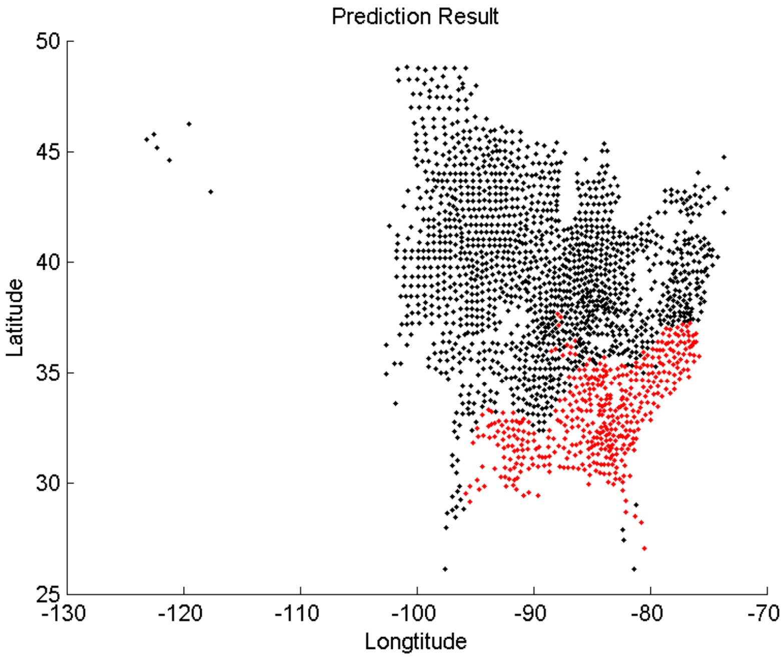

Figure 2.37: *Prediction for the last two weeks of October.*

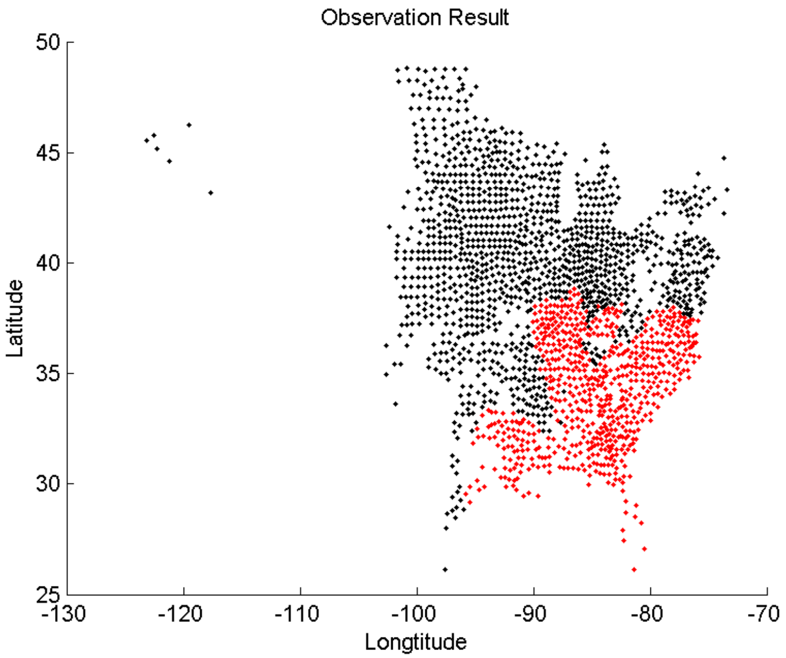

Figure 2.38: *Observations in the last two weeks of November.*

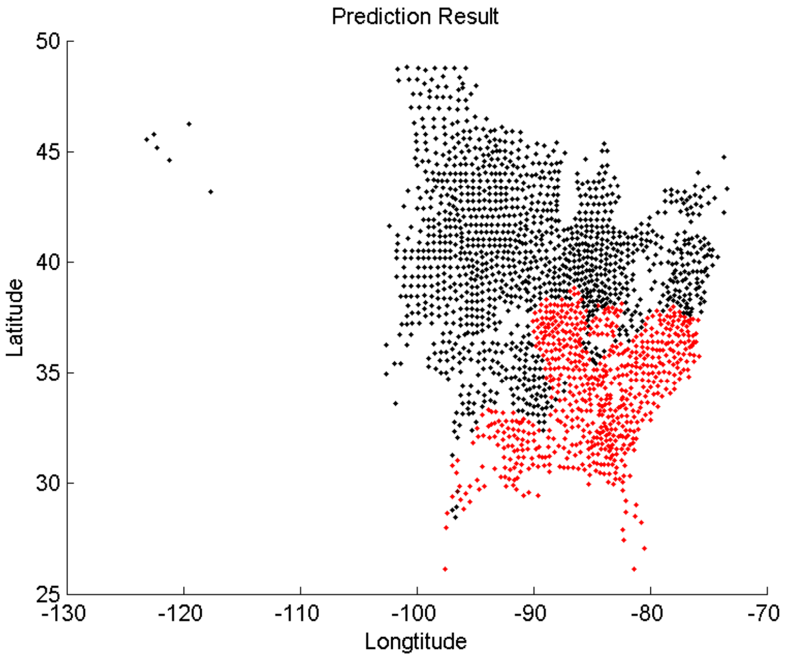

Figure 2.39: *Prediction for the last two weeks of November.*

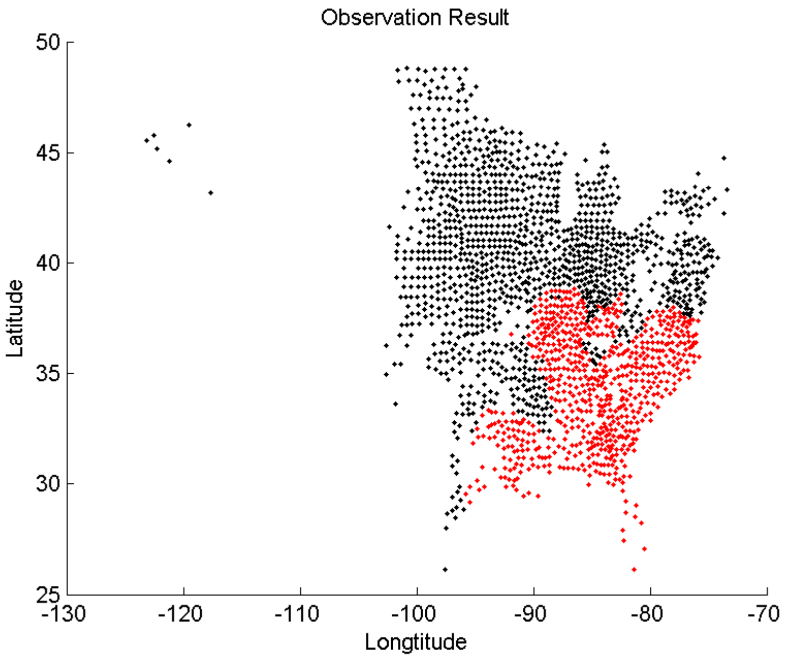

Figure 2.40: *Observations in the last two weeks of December.*

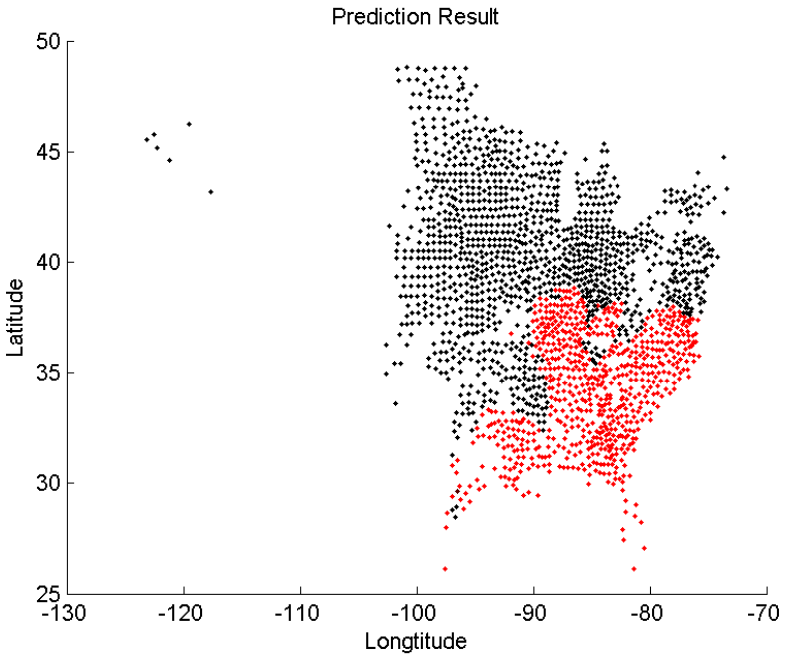

Figure 2.41: *Prediction for the last two weeks of December.*

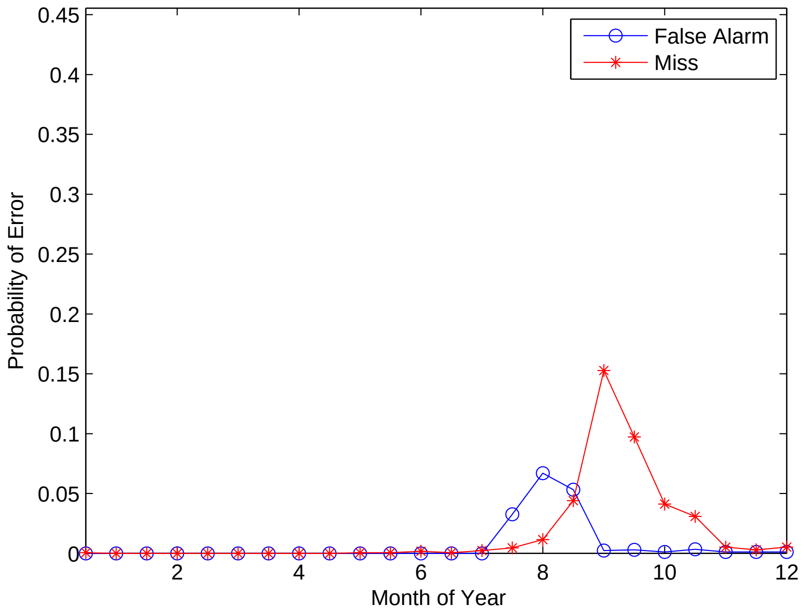

Figure 2.42: *Errors prediction for 2009 based on observations from 2005.*

2.3.1 Prediction Error

There are two different types of prediction error: false alarm error and miss error. False alarm error happens when the observation is healthy but the prediction of network is infected. If the observation is infected and the model's prediction is healthy, it is miss error. False alarm error at each time step can be calculated

$$E_{FA}(n) = \frac{1}{N} \sum_{j \in H_n} \pi_{j1}(n), \tag{2.24}$$

where N denotes the total number of counties, H_n denotes the set of counties whose observations are healthy at time step n, $\pi_{j1}(n)$ denotes the probability of infection for county j at time step n.
Miss error at each time step can be written

$$E_M(n) = \frac{1}{N} \sum_{j \in I_n} \{1 - \pi_{j1}(n)\}, \tag{2.25}$$

where I_n denotes the set of counties whose observations are infected at time step n.
(2.24) and (2.25) give us false alarm and miss errors at each time step respectively. Here for instance, we show the miss and false alarm errors over time for 2009's prediction in figures from 2.42 up to 2.56. In each figure, we consider different permutations of the rest of years to build the network. The corresponding error figures for the other years are shown in appendix A.

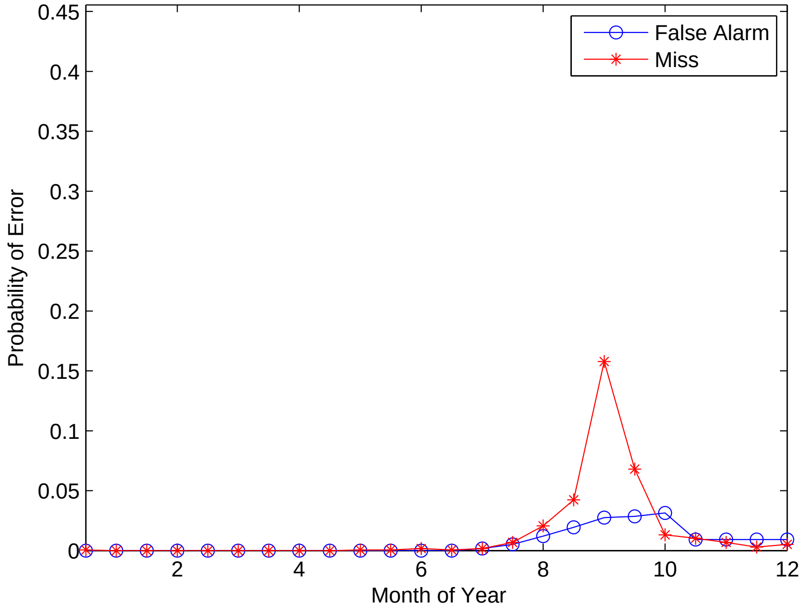

Figure 2.43: *Errors prediction for 2009 based on observations from 2006.*

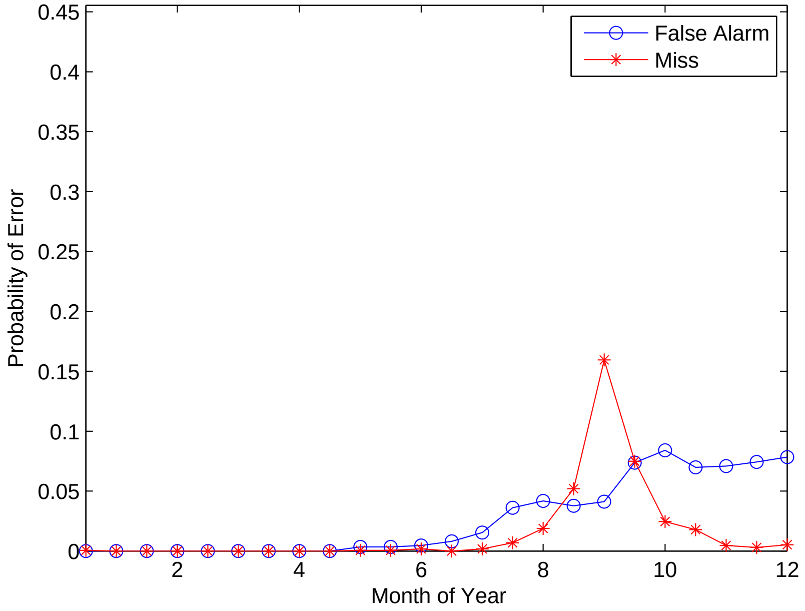

Figure 2.44: *Errors prediction for 2009 based on observations from 2007.*

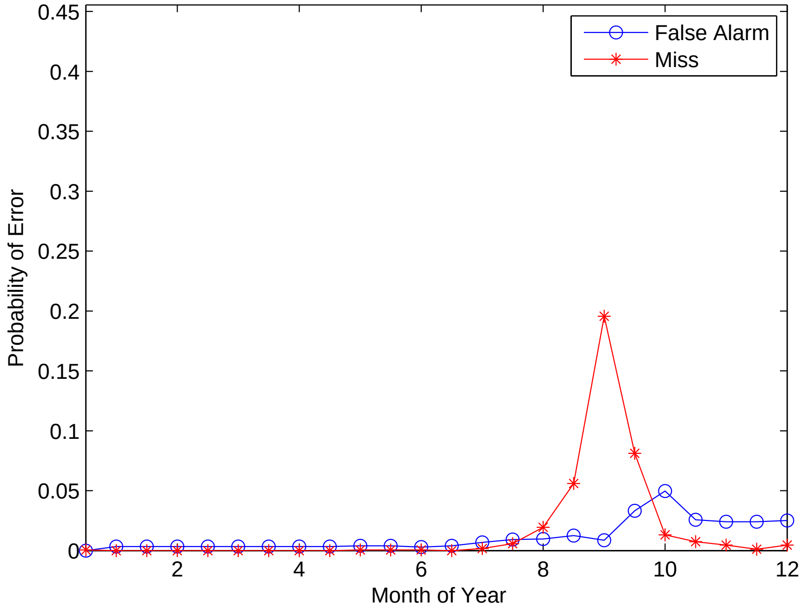

Figure 2.45: *Errors prediction for 2009 based on observations from 2008.*

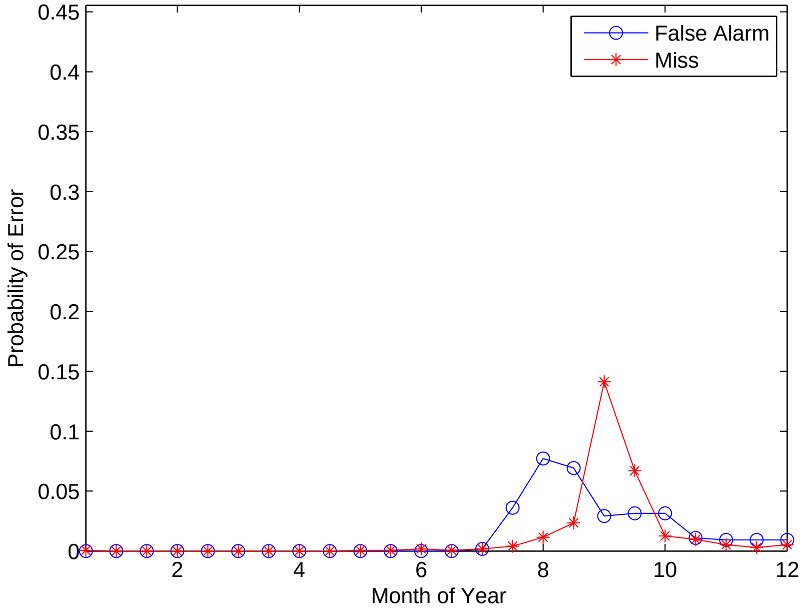

Figure 2.46: *Errors prediction for 2009 based on observations from 2005 and 2006.*

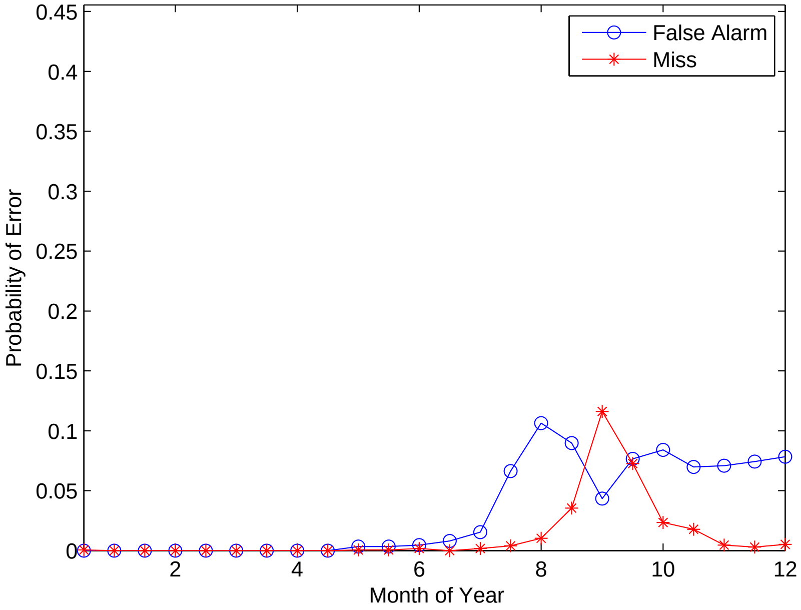

Figure 2.47: *Errors prediction for 2009 based on observations from 2005 and 2007.*

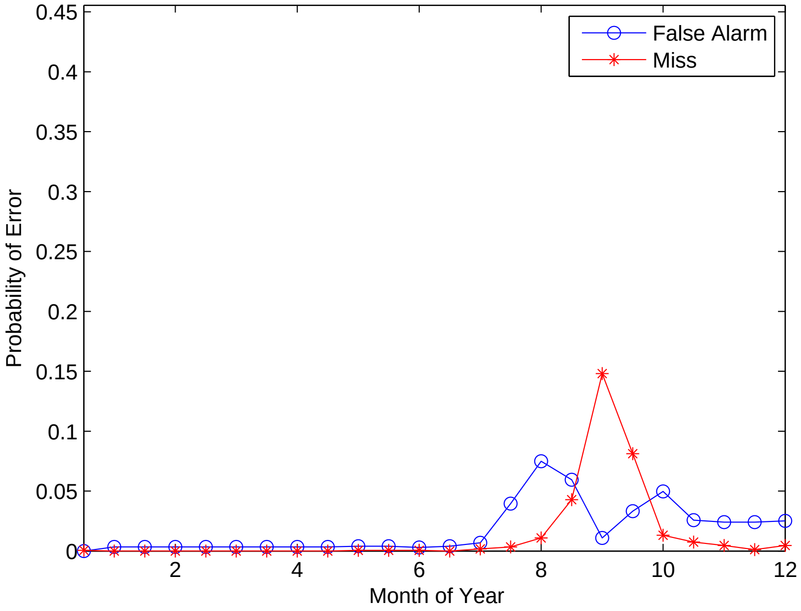

Figure 2.48: *Errors prediction for 2009 based on observations from 2005 and 2008.*

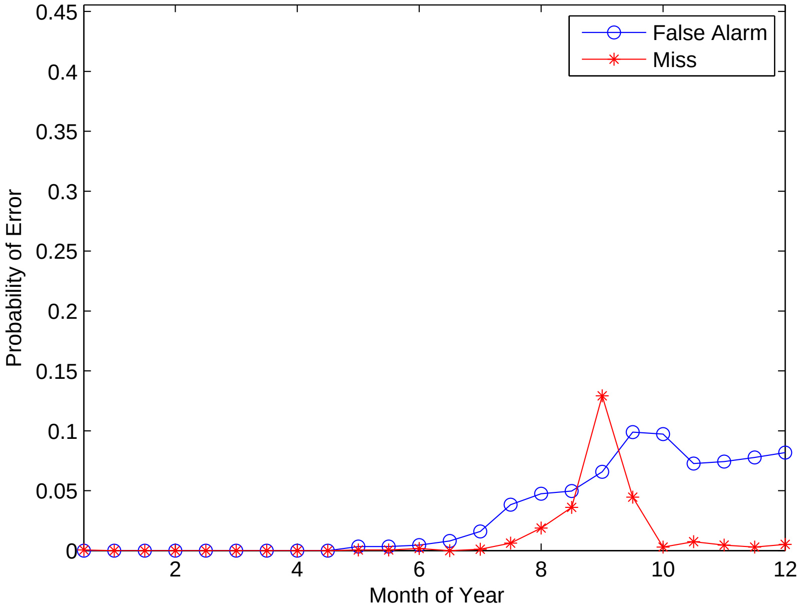

Figure 2.49: *Errors prediction for 2009 based on observations from 2006 and 2007.*

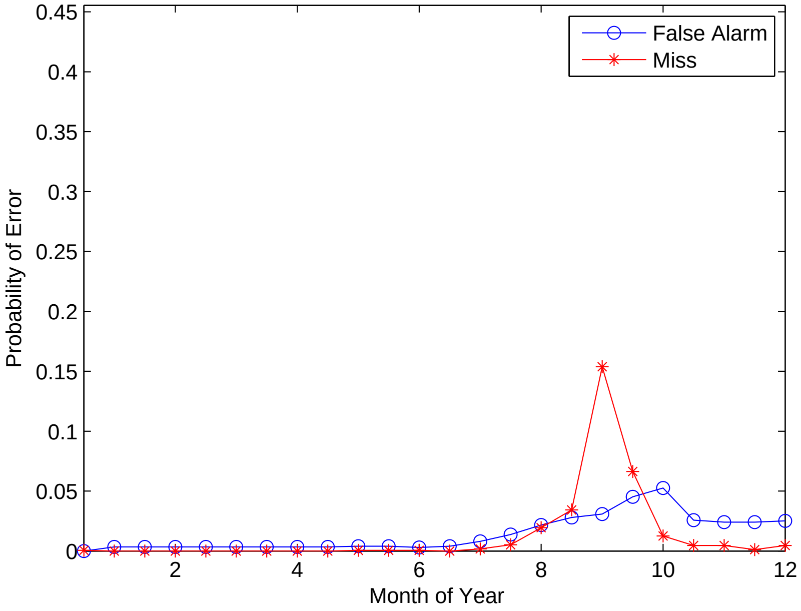

Figure 2.50: *Errors prediction for 2009 based on observations from 2006 and 2008.*

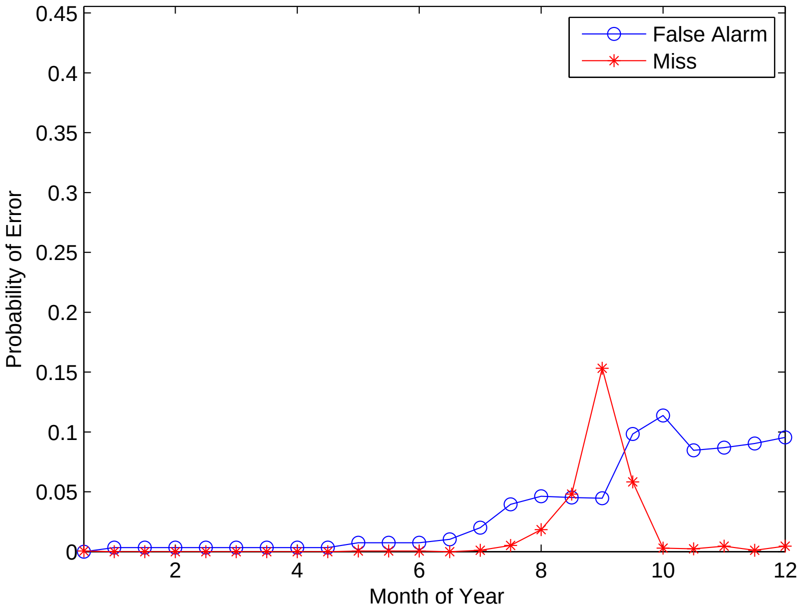

Figure 2.51: *Errors prediction for 2009 based on observations from 2007 and 2008.*

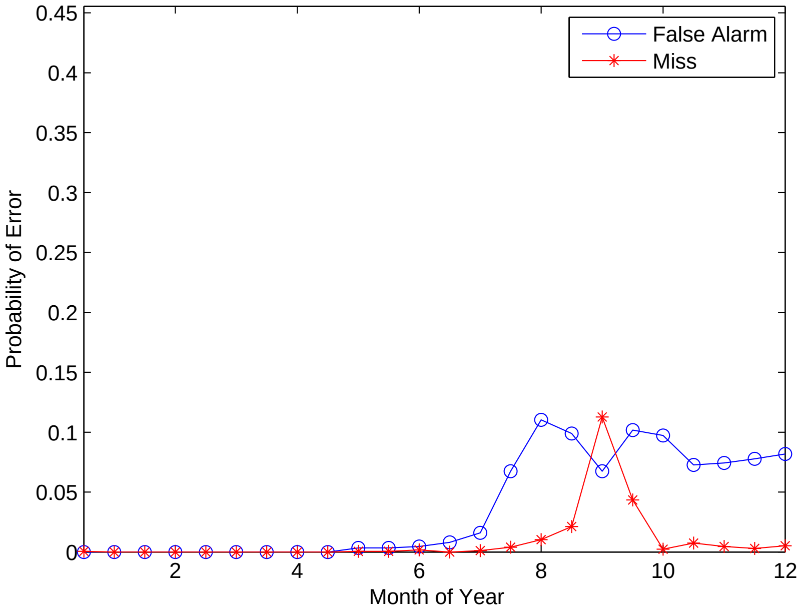

Figure 2.52: *Errors prediction for 2009 based on observations from 2005, 2006, and 2007.*

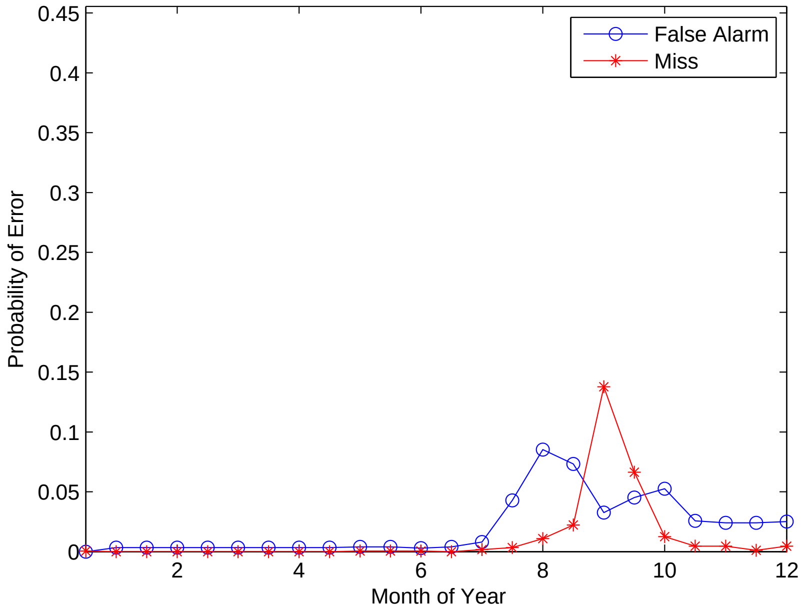

Figure 2.53: *Errors prediction for 2009 based on observations from 2005, 2006, and 2008.*

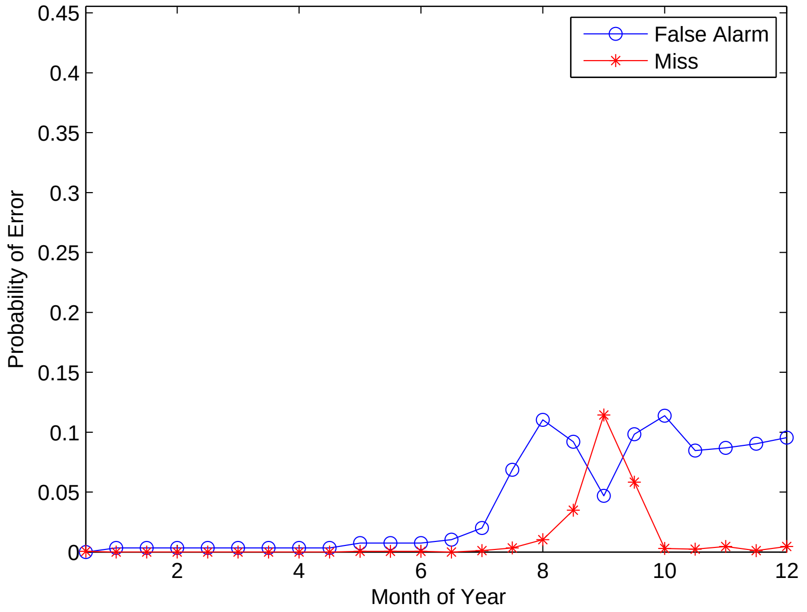

Figure 2.54: *Errors prediction for 2009 based on observations from 2005, 2007, and 2008.*

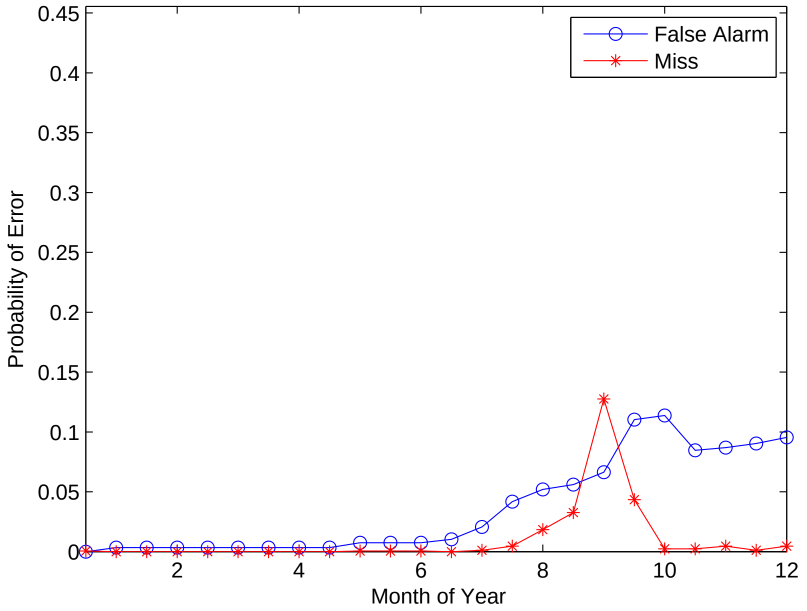

Figure 2.55: *Errors prediction for 2009 based on observations from 2006, 2007, and 2008.*

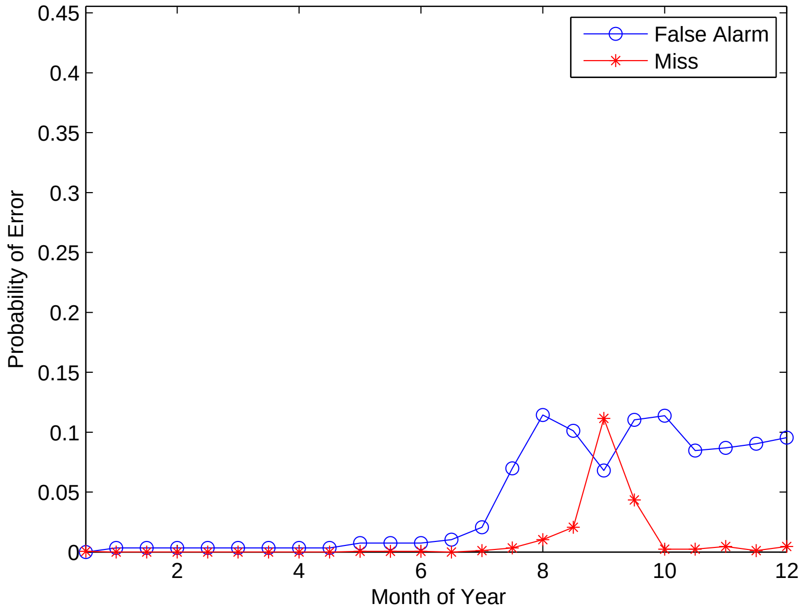

Figure 2.56: *Errors prediction for 2009 based on observations from 4 other years.*

The average miss and false alarm errors can be computed by averaging the miss and false alarm errors over time. The average false alarm and miss errors can be written respectively

$$\bar{E}_{FA} = \frac{1}{T} \sum_{n=T_0}^{T_{final}} E_{FA}(n), \tag{2.26}$$

$$\bar{E}_{M} = \frac{1}{T} \sum_{n=T_0}^{T_{final}} E_{M}(n), \tag{2.27}$$

where T_0 denotes the first time step at which there is an infected observation and T_{final} denotes the last time step which is the last two weeks of December in our study.

You can see the average false alarm and miss errors for the year 2005, 2006, 2007, 2008, and 2009 in tables from 2.1 up to 2.10. In these tables, errors for different cases of permutations of years as data to be used to estimate the network's parameters are shown. As it can be observed from these tables, the main trend for all the years is that increasing number of years which are used to build the network results in increasing false alarm error and decreasing miss error. This trend is reasonable because more years that are used to construct the model yields to the network which is more general and encompasses more different possible paths of disease dispersal. Considering more possible pattern of disease dispersal in previous years decreases the chance of missing one potential infected events. On the other hand, it increase the chance of false prediction of infected events for some counties which are not actually infected.

Table 2.1: *False Alarm Probability of 2005 Prediction*

Years	False Alarm Probability
2006	0.103304637600666
2007	0.150196374023089
2008	0.136271670567699
2009	0.181774903796564
2006 & 2007	0.195739278771770
2006 & 2008	0.155869401356766
2006 & 2009	0.194311104058396
2007 & 2008	0.202166064981949
2007 & 2009	0.230848573808863
2008 & 2009	0.200142817471337
2006 & 2007 & 2008	0.216011425397707
2006 & 2007 & 2009	0.237235688499226
2006 & 2008 & 2009	0.205101757448328
2007 & 2008 & 2009	0.241678898718610
2006 & 2007 & 2008 & 2009	0.245130320942595

Table 2.2: *Miss Probability of 2005 Prediction*

Years	Miss Probability
2006	0.00813266156226445
2007	0.00813266156226445
2008	0.00725988812631412
2009	0.00658547228944341
2006 & 2007	0.00761693180465744
2006 & 2008	0.00710120204705042
2006 & 2009	0.00634744317054786
2007 & 2008	0.00710120204705042
2007 & 2009	0.00622842861110009
2008 & 2009	0.00587138493275677
2006 & 2007 & 2008	0.00694251596778673
2006 & 2007 & 2009	0.00610941405165232
2006 & 2008 & 2009	0.00579204189312493
2007 & 2008 & 2009	0.00579204189312493
2006 & 2007 & 2008 & 2009	0.00571269885349308

Table 2.3: *False Alarm Probability of 2006 Prediction*

Years	False Alarm Probability
2005	0.0194285714285714
2007	0.0831861471861472
2008	0.0527445887445888
2009	0.0887965367965368
2005 & 2007	0.0992207792207792
2005 & 2008	0.0656277056277056
2005 & 2009	0.0995197026397947
2007 & 2008	0.104138528138528
2007 & 2009	0.126428793548886
2008 & 2009	0.0993465424666346
2005 & 2007 & 2008	0.116709956709957
2005 & 2007 & 2009	0.136672586752648
2005 & 2008 & 2009	0.109140119220181
2007 & 2008 & 2009	0.133521071601133
2005 & 2007 & 2008 & 2009	0.143145132705179

Table 2.4: *Miss Probability of 2006 Prediction*

Years	Miss Probability
2005	0.0164848484848485
2007	0.0158268398268398
2008	0.00976623376623377
2009	0.00550649350649351
2005 & 2007	0.0122597402597403
2005 & 2008	0.00737662337662338
2005 & 2009	0.00516017316017316
2007 & 2008	0.00817316017316017
2007 & 2009	0.00387878787878788
2008 & 2009	0.00335930735930736
2005 & 2007 & 2008	0.00633766233766234
2005 & 2007 & 2009	0.00363636363636364
2005 & 2008 & 2009	0.00315151515151515
2007 & 2008 & 2009	0.00263203463203463
2005 & 2007 & 2008 & 2009	0.00242424242424242

Table 2.5: *False Alarm Probability of 2007 Prediction*

Years	False Alarm Probability
2005	0.0160239242992323
2006	0.0312667380824853
2008	0.0377164792001428
2009	0.0546107837886092
2005 & 2006	0.0415550794500982
2005 & 2008	0.0471121228352080
2005 & 2009	0.0614176039992858
2006 & 2008	0.0470005356186395
2006 & 2009	0.0585386538118193
2008 & 2009	0.0629351901446170
2005 & 2006 & 2008	0.0549455454383146
2005 & 2006 & 2009	0.0650776647027317
2005 & 2008 & 2009	0.0692287091590787
2006 & 2008 & 2009	0.0650553472594180
2005 & 2006 & 2008 & 2009	0.0711256918407427

Table 2.6: *Miss Probability of 2007 Prediction*

Years	Miss Probability
2005	0.0131003392251384
2006	0.00883770755222282
2008	0.00836904124263524
2009	0.00651669344759864
2005 & 2006	0.00819050169612569
2005 & 2008	0.00794500981967506
2005 & 2009	0.00638278878771648
2006 & 2008	0.00694072487055883
2006 & 2009	0.00584717014818782
2008 & 2009	0.00578021781824674
2005 & 2006 & 2008	0.00687377254061775
2005 & 2006 & 2009	0.00578021781824674
2005 & 2008 & 2009	0.00578021781824674
2006 & 2008 & 2009	0.00557936082842350
2005 & 2006 & 2008 & 2009	0.00557936082842350

Table 2.7: *False Alarm Probability of 2008 Prediction*

Years	False Alarm Probability
2005	0.0105233219567691
2006	0.0134859689040576
2007	0.0368174742516695
2009	0.0402445961319681
2005 & 2006	0.0220894956389837
2005 & 2007	0.0466184478427961
2005 & 2009	0.0469283276450512
2006 & 2007	0.0465236431404829
2006 & 2009	0.0431124383769435
2007 & 2009	0.0643232259997927
2005 & 2006 & 2008	0.0548907462928716
2005 & 2006 & 2009	0.0495828593098218
2005 & 2007 & 2009	0.0708653386326516
2006 & 2007 & 2009	0.0664806211506645
2005 & 2006 & 2007 & 2009	0.0728209797044849

Table 2.8: *Miss Probability of 2008 Prediction*

Years	Miss Probability
2005	0.014031095942359
2006	0.007916192643155
2007	0.009907091391733
2009	0.003531475161168
2005 & 2006	0.006257110352673
2005 & 2007	0.007631778536215
2005 & 2009	0.003104854000758
2006 & 2008	0.004479522184300
2006 & 2009	0.002156806977626
2007 & 2009	0.002986348122867
2005 & 2006 & 2007	0.003104854000758
2005 & 2006 & 2009	0.001801289343951
2005 & 2007 & 2009	0.002607129313614
2006 & 2007 & 2009	0.001919795221843
2005 & 2006 & 2007 & 2009	0.001587978763747

Table 2.9: *False Alarm Probability of 2009 Prediction*

Years	False Alarm Probability
2005	0.006908709738898
2006	0.006765770916714
2007	0.026777206022489
2008	0.011292166952544
2005 & 2006	0.013102725366876
2005 & 2007	0.033114160472651
2005 & 2008	0.017319420621307
2006 & 2007	0.030803316180675
2006 & 2008	0.014222412807318
2007 & 2008	0.034376786735277
2005 & 2006 & 2007	0.036878216123499
2005 & 2006 & 2008	0.020035258242805
2005 & 2007 & 2008	0.040308747855918
2006 & 2007 & 2008	0.036592338479131
2005 & 2006 & 2007 & 2008	0.042309891366495

Table 2.10: *Miss Probability of 2009 Prediction*

Years	Miss Probability
2005	0.016723842195540
2006	0.014150943396226
2007	0.015508862206975
2008	0.016366495140080
2005 & 2006	0.012006861063465
2005 & 2007	0.012388031255956
2005 & 2008	0.013388603011245
2006 & 2007	0.010934819897084
2006 & 2008	0.012935963407662
2007 & 2008	0.012578616352201
2005 & 2006 & 2007	0.009100438345721
2005 & 2006 & 2008	0.011339813226606
2005 & 2007 & 2008	0.010005717552887
2006 & 2007 & 2008	0.010196302649133
2005 & 2006 & 2007 & 2008	0.008647798742138

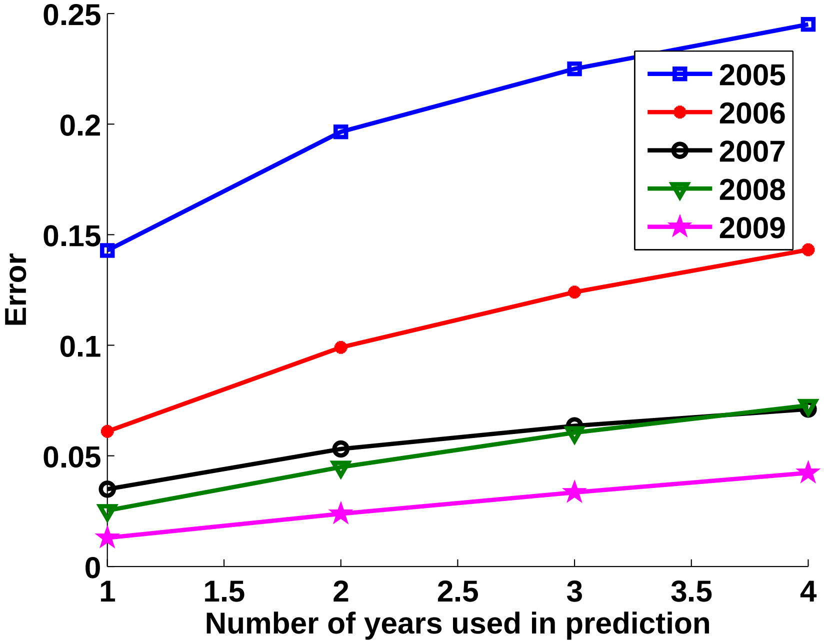

Figure 2.57: *False alarm probability of error versus number of years used for prediction.*

2.3.2 Average Prediction Error

The more years of data used for predicting results, the better the fit is?

First, we discuss about the hypothesis that the more years of data used to build the network results in more accurate prediction. Fig. 2.57 shows the average false alarm error for all the years versus the number of years of data. As you can see, in all the years the more years of data yields to increase in false alarm error because more years of data increases the chance of false prediction of infected event when there is no infection actually. Fig. 2.58 plot the average miss error versus the number of years of data. It can be observed that more number of years of data decreases the average miss error. More years of data makes the model more complete by providing it with more possible patterns of disease dispersal. Therefore, the hypothesis is right in the case of miss error and is wrong in the case of false alarm error.

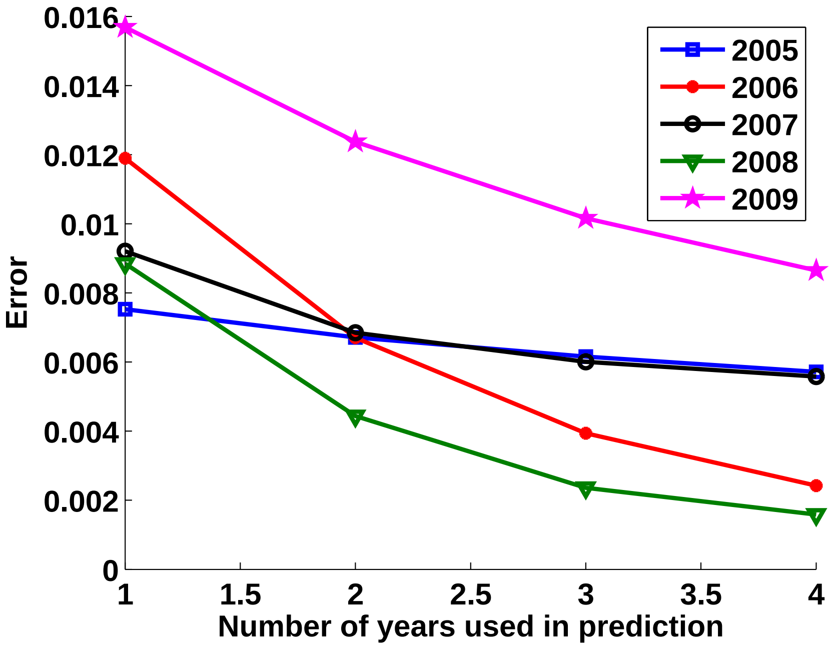

Figure 2.58: *Miss probability of error versus number of years used for prediction.*

The years closer to the year being predicted will provide better estimates

The second hypothesis which we discuss about is that the years closer to the year being predicted provide better estimates. Fig. 2.59 plots the average false alarm error versus distance between predicted and predictor years. As you can see, there is not any consistent decreasing or increasing behavior in these curves, which shows that the distance between predicted and predictor years does not matter in the amount of the average false alarm error. Fig. 2.60 shows the average miss alarm error versus distance between predicted and predictor years. As it can be seen, there is not any consistent decreasing or increasing behavior in the curves of miss error like false alarm curves, which shows that the distance between predicted and predictor years does not have any role in decrease of the average miss error so that the hypothesis is wrong.

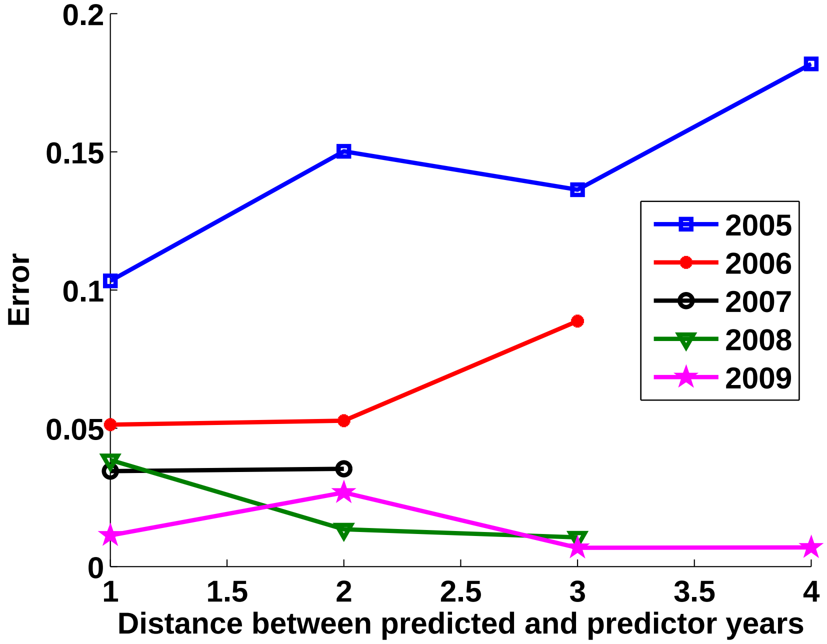

Figure 2.59: *False alarm probability of error versus distance between predicted and predictor years.*

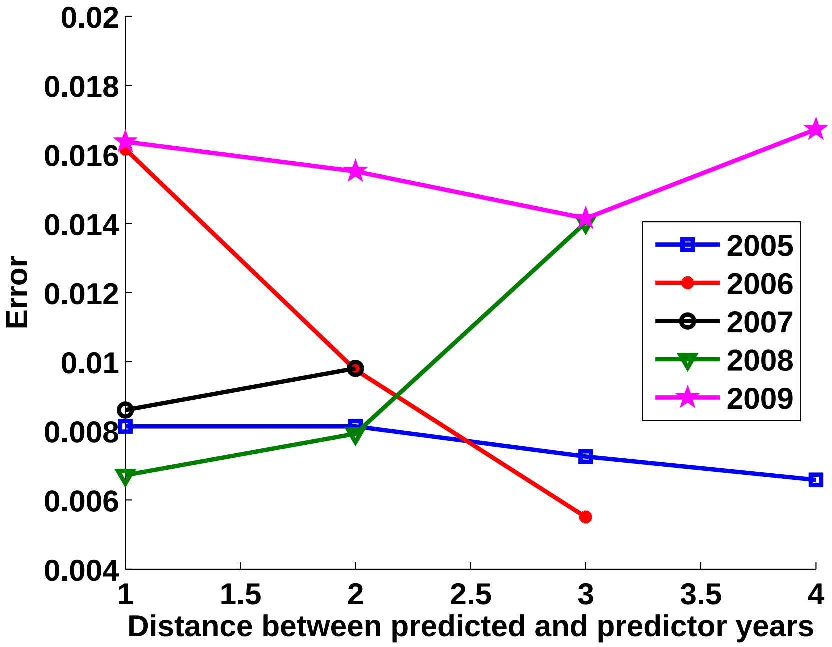

Figure 2.60: *Miss probability of error versus distance between predicted and predictor years.*

The years before the year being predicted will provide better estimates than the years after

The third hypothesis which we discuss about is that the years before the year being predicted provide better estimates. Fig. 2.61 shows the average false alarm error versus difference in time between predicted and predictor years. The dominant trend in this figure is that the years before the year being predicted result in smaller average false alarm error compare to the years after that. It means that in the case of false alarm error, the casuality makes scenes and the years before the year being predicted are the more relevant years to build the network. Therefore, the hypothesis is confirmed in the case of false alarm error. Fig. 2.62 plots the average miss alarm error versus difference in time between predicted and predictor years. This figure shows that there is no consistent decreasing or increasing behavior in the average miss error in terms of difference in time between predicted and predictor years, which does not confirm the hypothesis.

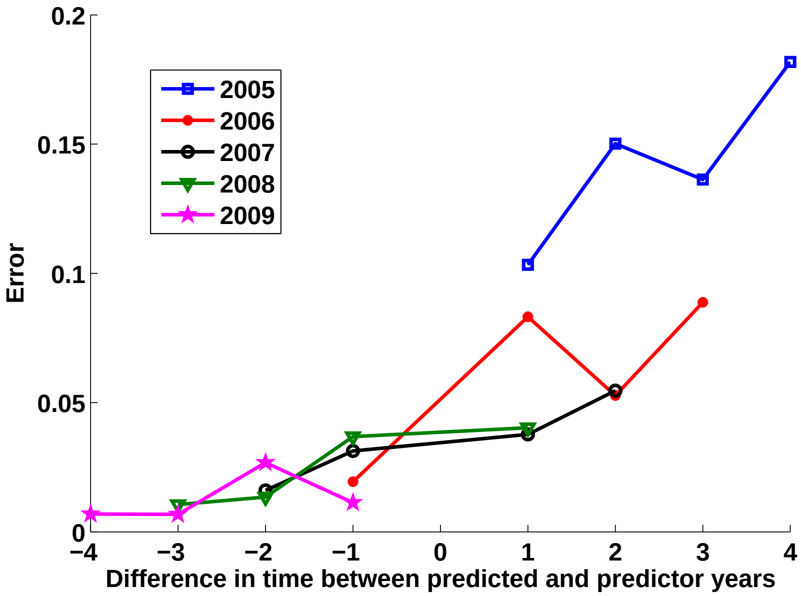

Figure 2.61: *False alarm probability of error versus difference in time between predicted and predictor years.*

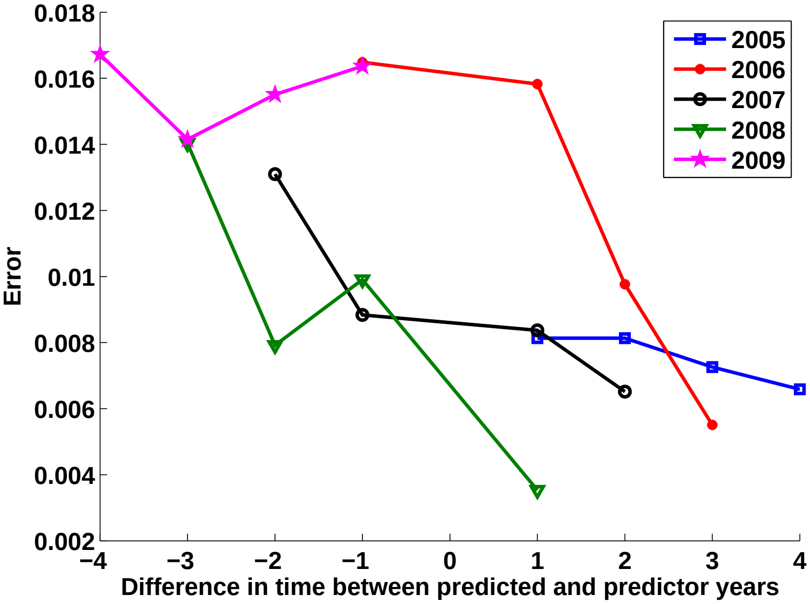

Figure 2.62: *Miss probability of error versus difference in time between predicted and predictor years.*

2.3.3 Comparison between prediction error early in the season with lately in the season

So far, we only discussed about the average false alarm and miss errors and we did not take into account the possible difference in behavior of these two errors early in the season or lately in the season. In our figures, time step corresponding to the end of the second week of July represents the early time in the season and time step corresponding to the end of the second week of October represents the lately time in the season.

The more years of data used for predicting results, the better the fit is?

In this section, we want to address this question: Is it more important to have more years of information early in the season than later in the season? Fig. 2.63 and Fig. 2.64 show false alarm error for all the years versus the number of years of data, respectively, early in the season and lately in the season. Although false alarm errors early in the season and lately in the season have the similar increasing trend in terms of the number of years data used, there are two differences between them. First, the amount of false alarm error lately in the season is more than false alarm error early in the season. Secondly, the greater number of years of data use for prediction causes higher increase in false alarm error early in than season rather than lately in the season. For example in 2005, false alarm error early in the season for the case of using four years data is about 3.5 times bigger than the false alarm error early in the season for the case of one year of data used for prediction. While, this increase for false alarm error lately in the season is about 2 times.
Fig. 2.65 and Fig. 2.66 plot miss error versus the number of years of data, respectively, early in the season and lately in the season. In the both figures, you can see that miss error decreases by increasing the number of years of data used to build model. However, the figures of miss error lately in the season decrease more rapidly compare to the ones early in the season.

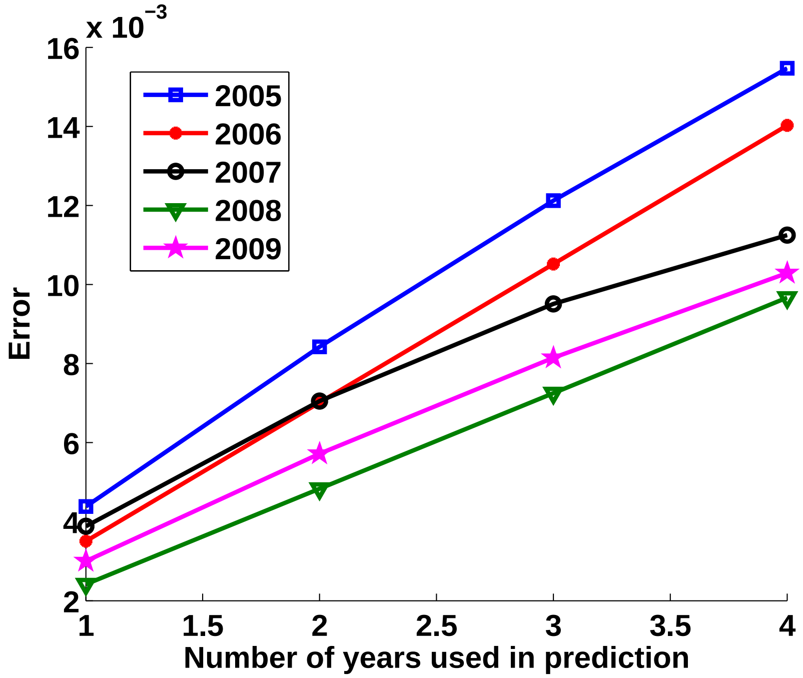

Figure 2.63: *False alarm probability of error versus number of years used for prediction in the second week of July.*

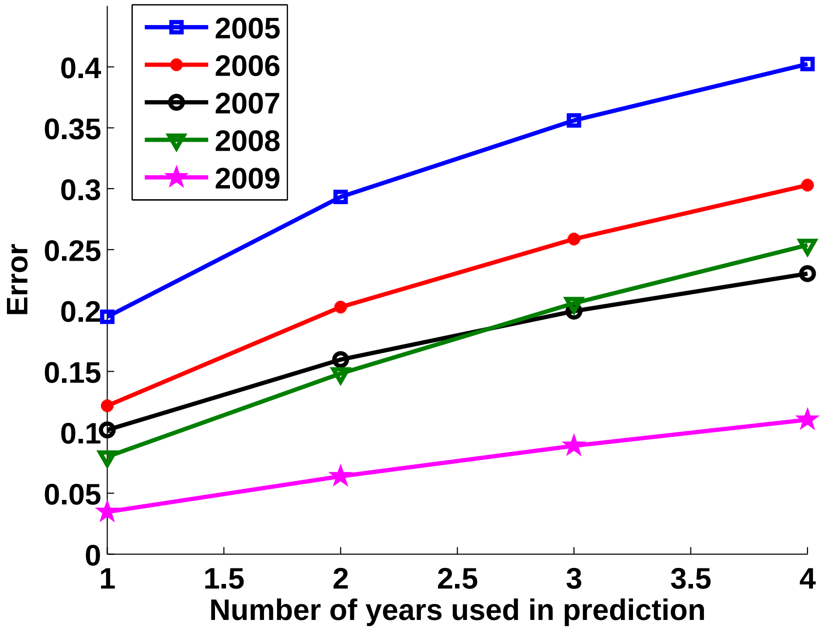

Figure 2.64: *False alarm probability of error versus number of years used for prediction in the second week of October.*

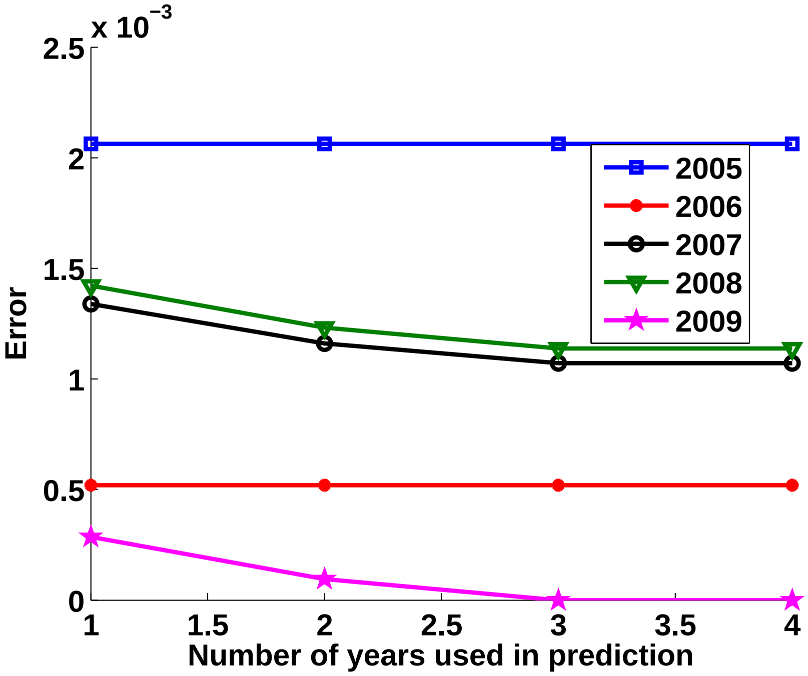

Figure 2.65: *Miss probability of error versus number of years used for prediction in the second week of July.*

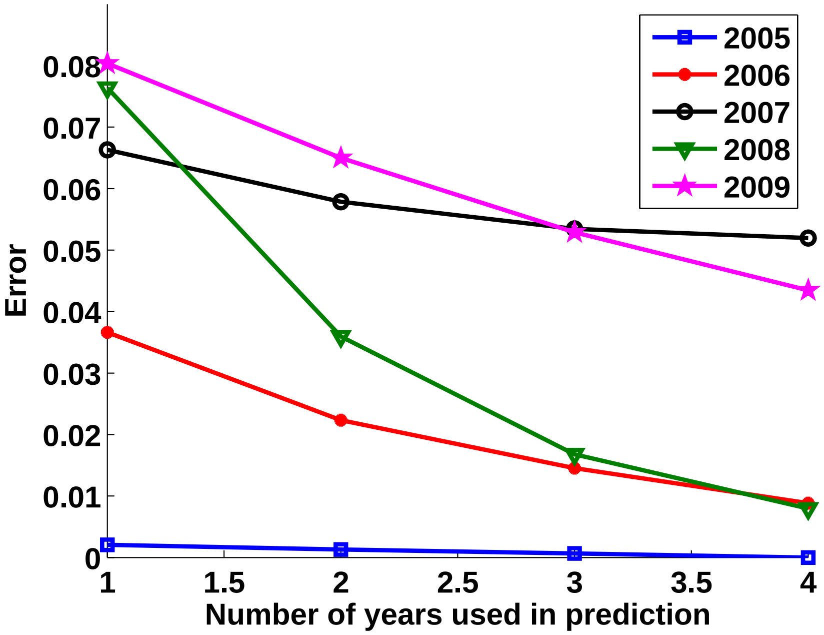

Figure 2.66: *Miss probability of error versus number of years used for prediction in the second week of October.*

The years closer to the year being predicted will provide better estimates

Fig. 2.69 and Fig. 2.70 plot false alarm error, respectively, early in the season and lately in the season versus distance between predicted and predictor years. Fig. 2.69 and Fig. 2.70 show miss error, respectively, early in the season and lately in the season versus distance between predicted and predictor years. As you can see in these four figures, there is no consistent decreasing or increasing behavior in any of these figures, which shows that the distance between predicted and predictor years does not matter in the amount of false alarm and miss errors early in the season or lately in the season.

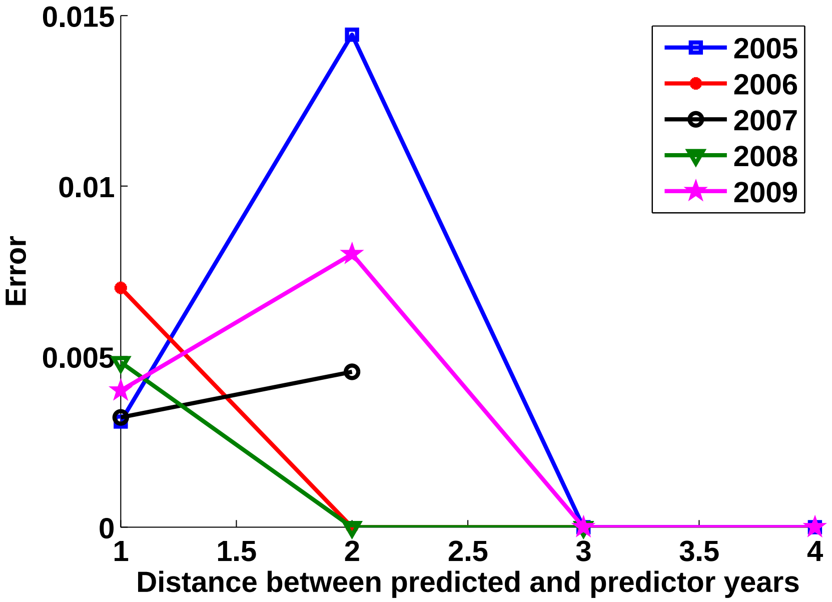

Figure 2.67: *False alarm probability of error versus distance between predicted and predictor years in the second week of July.*

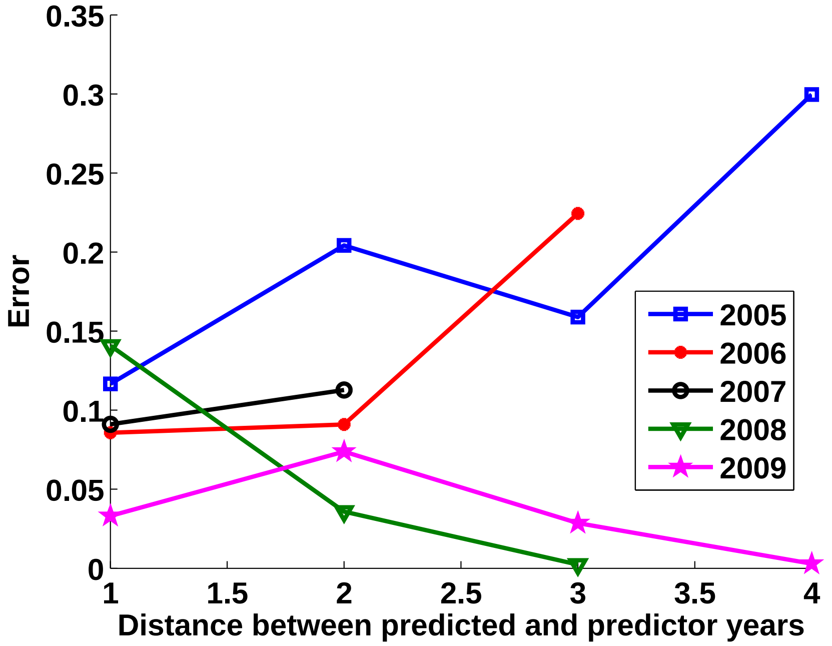

Figure 2.68: *False alarm probability of error versus distance between predicted and predictor years in the second week of October.*

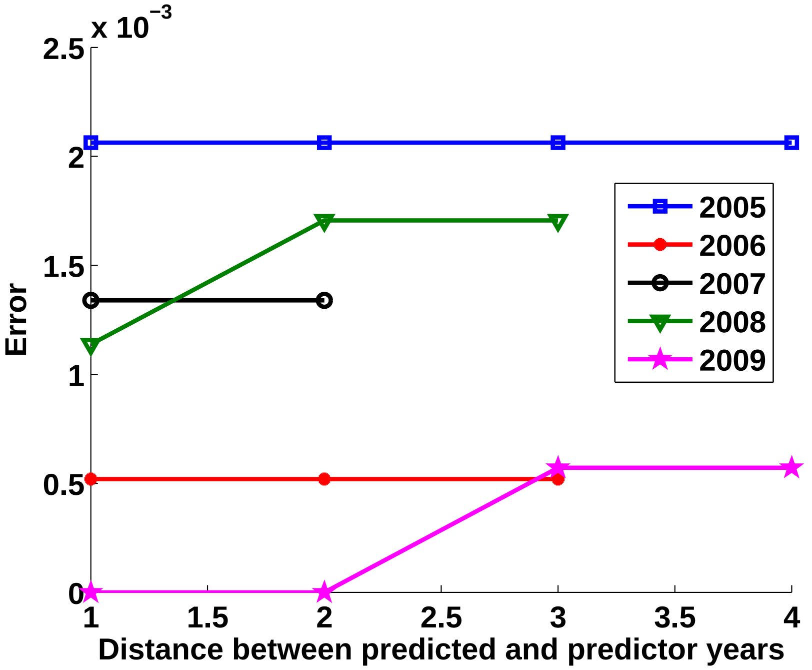

Figure 2.69: *Miss probability of error versus distance between predicted and predictor years in the second week of July.*

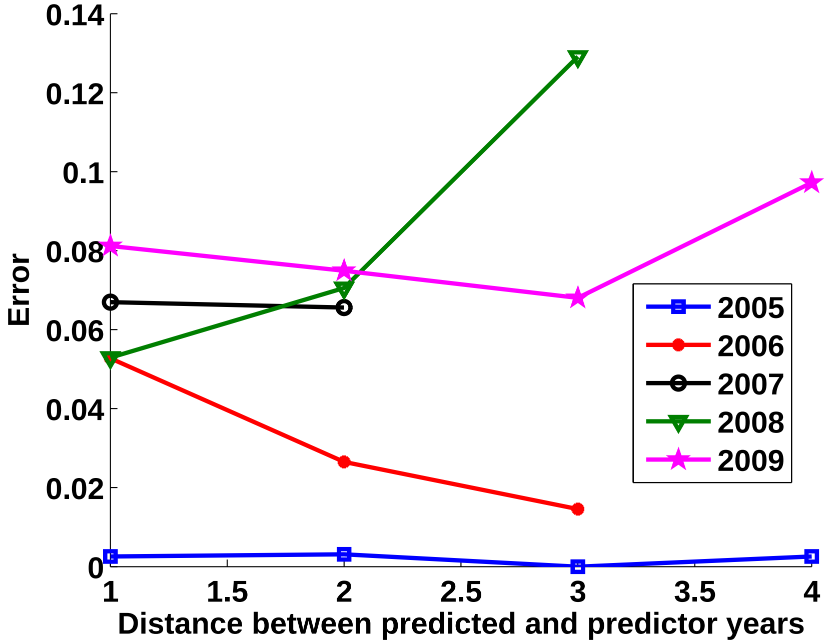

Figure 2.70: *Miss probability of error versus distance between predicted and predictor years in the second week of October.*

The years before the year being predicted will provide better estimates than the years after

Fig. 2.71 shows false alarm error early in the season versus difference in time between predicted and predictor years. As you can see, there is no consistent trend in the curves of false alarm error early in the season. Fig. 2.72 plots false alarm error lately in the season versus difference in time between predicted and predictor years. It can be seen that the years before the year being predicted result in smaller false alarm error compare to the years after that. Fig. 2.73 and Fig. 2.74 plot miss alarm error, respectively, early and lately in the season versus difference in time between predicted and predictor years. The interesting observation in these two figures is that miss error is smaller when we use the years' data after the year being predicted compare to the case when the years before the year being predicted are used to construct the model.

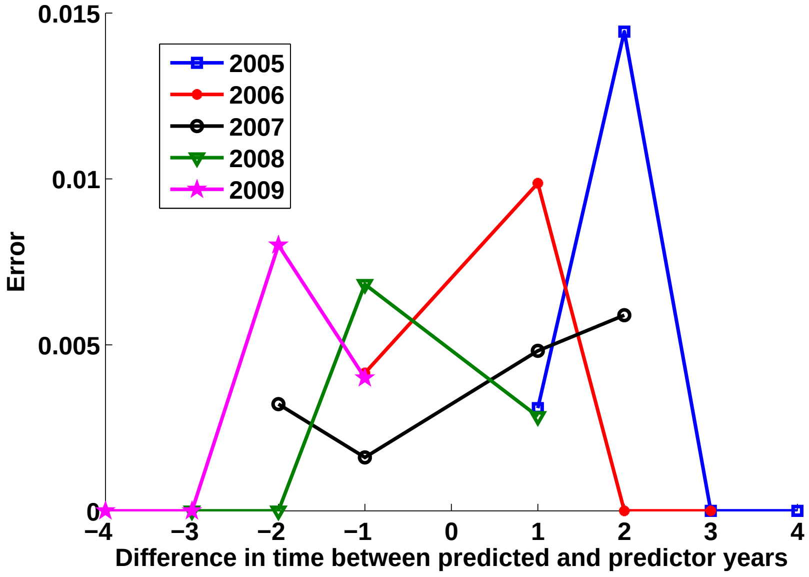

Figure 2.71: *False alarm probability of error versus difference between predicted and predictor years in the second week of July.*

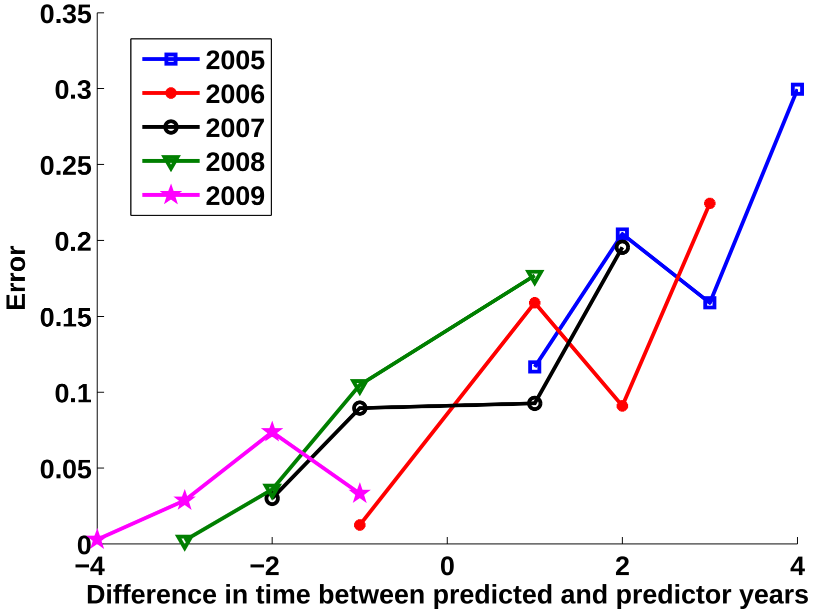

Figure 2.72: *False alarm probability of error versus difference between predicted and predictor years in the second week of October.*

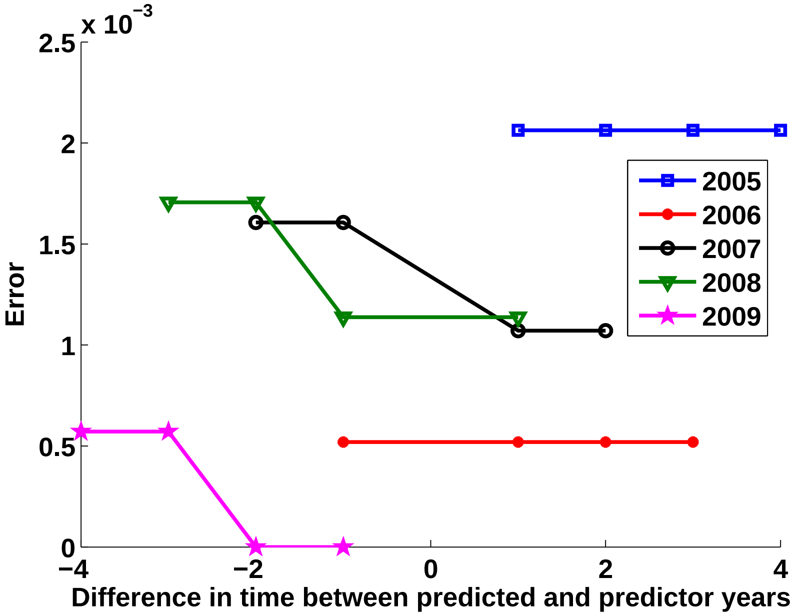

Figure 2.73: *Miss probability of error versus difference between predicted and predictor years in the second week of July.*

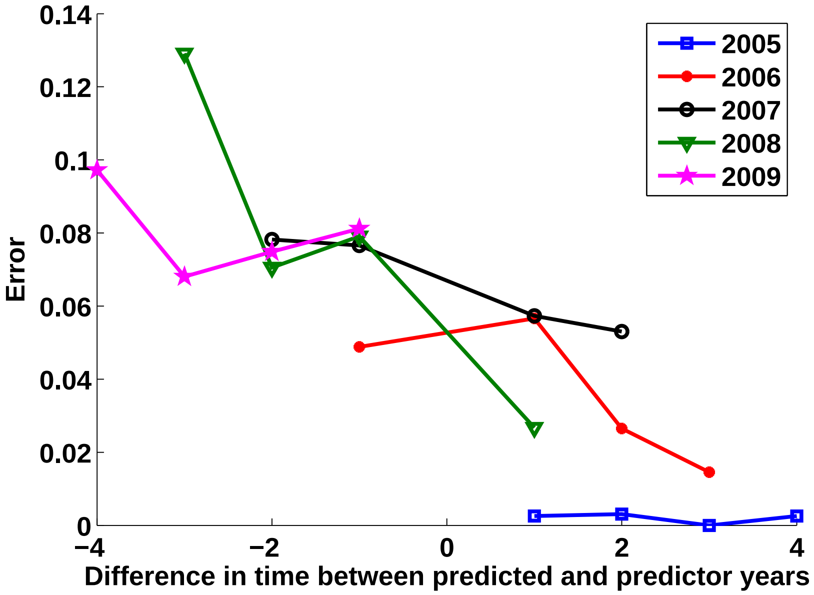

Figure 2.74: *Miss probability of error versus difference between predicted and predictor years in the second week of October.*

Chapter 3

Malware Propagation Modeling with Spatial Correlation on Real Networks

3.1 Introduction

The rapid growth in the number of Internet users in recent years has led malware generators to exploit this potential to attack computer users around the word. Internet users are frequent targets of malicious software every day. The ability of malware to exploit the infrastructures of networks for propagation determines how detrimental they can be to the network's security. Malicious software can make large outbreaks if they are able to exploit the structure of the Internet and interactions between users to propagate.

Malware can be classified based on two main propagation schemes: random and topological scanning. While random scanning viruses use random scanning of IP addresses to find victims, topological worms employ the social networking facets of different services on the Internet, like email contact lists and online social networking sites such as facebook and twitter to find the potential neighbors. Characterizing various malware propagation schemes can be done by choosing appropriate topologies as underlying graphs. The topologies of networks have the significant effects on the propagation of malwares. For example, scale free networks are more vulnerable to malwares rather random networks because of having hops.

For random scanning types of malware, the network topologies can be modeled by Erdos-Renyi (ER) graphs. For topological malware, underlying topologies are specific overlay topologies created by applications. The underlying topology for a topological malware like Koobface, which uses the friend list of each user in facebook as future potential targets to attack, is the topology of friendship contacts on facebook[14].

In epidemic networks, nodes can get infected through their infected neighbors. If a healthy node does not have any infected neighbors, it will be healthy forever. SIR (Susceptible-Infected-Removed) and SIS (Susceptible-Infected-Susceptible) are two primary epidemic models in the epidemic literature. In the SIR model, an infected node is not susceptible to malware anymore once it has been cured. In the SIS model, an infected node's state after being cured is susceptible again. Ganesh et al. proposed conditions related to the topologies of networks for fast and slow dying out of epidemics[8]. They characterize the life time of epidemics by comparing the epidemic strength with the spectral radius of graph and the isoperimetric constant of graph. In[3], Chakrabarti et al. proposed a nonlinear dynamical system in discrete time based on the assumption of independence of nodes' states to model viral propagation in a network. Wang et al. developed a discrete time model using the assumption of independence to estimate epidemic spreading and to obtain an epidemic threshold[17]. In[10], Van Mieghem et al. proposed the N-intertwined model based on the assumption of independence and the exact 2^N state Markov chain model in continuous time to model virus propagation. Chen et al. presented a Markov model which incorporates partial spatial dependence between neighbors' states in discrete time and showed that this model outperforms the independent model[4]. In the context of discrete time modeling for epidemic, it is assumed that just one event is allowed at each one single time step[4] or more than one event, which considers concurrence of infection and recovery during one time step[3,17]. It is remarkable that these two different assumptions yield to variant transient behaviors of epidemic and also distinct values for final number of infected nodes. We should notice that in the all mentioned references, it is assumed that the cure probability is independent of nodes' states.

In this chapter, we study the SIS malware spreading in discrete time while allowing one event at each single discrete time step. We assume that the cure probability at each node depends on its neighbours' states. To the best of our knowledge, we are the first to propose analytical modeling for the propagation of malware with considering spatial dependence between neighbors' states while assuming the dependence of cure probability to nodes' states. First, we develop the full model which considers the exact spatial dependence between neighbors' states without any approximation. Because obtaining the exact spatial dependence formula is computationally too expensive, we develop the independent and the Markov models to approximate it. In the independent model, although we assume spatial independence between neighbors' states, the detailed network topology and the temporal dependence are considered to develop the model. In the Markov model, unlike the independent model, we take into account the partial spatial dependence between neighbors' states when computing the two-node joint probability to calculate the conditional infection probability of each node. Finally, we compare results of the Markov and the independent models with the Monte Carlo simulation results for different real and synthesized topologies. We show that the topologies of networks or, in this case malware scanning schemes, have significant effects on the performance of the Markov and independent models. In all cases, we show that the Markov model outperforms the independent model. In particular, these two models' results closely overlap with the simulation in the steady state, while the Markov model yields far better performance than the independent model in the transient state in most cases. The initial fractions of infected nodes affect the transient behaviors of epidemic dispersal. We show that the gain of employing the Markov model instead of the independent model is more significant if the initial fraction of infected nodes is small. Increasing the connectivity level of networks makes the growth of epidemics faster in networks; as a consequence the Markov model yields far better results than the independent model when networks have a low level of connection. We show that the advantage of exploiting the Markov model rather than the independent model is more substantial if epidemic strength is low.

The rest of the chapter is organized as follows. In section 3.2, the full model is introduced. In section 3.3, we develop the independent model. In section 3.4, an analytical method to derive the Markov model is proposed. Section 4.4 contains simulation results that validate our theoretical analysis.

3.2 FULL Model

The correlation between neighbors' states increases over time with epidemic evolution. In a network, the state of each node is spatially dependent on its neighbors' states. At the same time, its neighbors' states are spatially dependent on their own neighbors' states. It causes that every node's state to be spatially dependent on all other nodes' states in the network. State of node i at time step n is represented by $X_i(n)$. $X_i(n)$ can be 0 (healthy) or 1 (infected). We assume that if node j is infected and neighbor of healthy node i, it contaminates node i with the probability of β_{ji}. Also, we assume that the cure probability is δ for all nodes when they do not have infected neighbours. We assume that having infected neighbours decreases the cure probability of infected nodes. Therefore, the transition probability from state 1 to state 0 is the joint probability of cure and not getting infection from infected neighbors and is equal to $\delta \prod_{j \in N_i} (1-\beta_{ji})^{x_j(n)}$ where N_i is a set of neighbors of node i and $x_j(n)$ is realization of node j's state at time step n. These assumptions define a discrete time Markov process. For specific realization of node i's neighbors' states at time step n, its Markov's transition probabilities can be written as

$$\begin{aligned} X_i &: 0 \to 1 \text{ with probability } 1 - \prod_{j \in N_i} (1-\beta_{ji})^{x_j(n)} \\ X_i &: 1 \to 0 \text{ with probability } \delta \prod_{j \in N_i} (1-\beta_{ji})^{x_j(n)}. \end{aligned} \tag{3.1}$$

Hence, transition probability matrix of node i at time step n can be written

$$\pi_i(n) = \begin{bmatrix} 1 - P_{i_{01}}(n) & P_{i_{01}}(n) \\ P_{i_{10}}(n) & 1 - P_{i_{10}}(n) \end{bmatrix}, \tag{3.2}$$

where $P_{i_{01}}(n)$ is the conditional probability that given node i is healthy at time step n, it will become infected at next time step; $P_{i_{10}}(n)$ is the conditional probability that given node i is infected at time step n, it will become healthy at next time step. The probability vector of node i's state at time step n can be written as

$$P_i(n) = \begin{bmatrix} P_{i_0}(n) & P_{i_1}(n) \end{bmatrix}, \tag{3.3}$$

where $P_{i_0}(n)$ is the probability that node i is healthy at time step n; $P_{i_1}(n)$ is the probability that node i is infected at time step n. The probability vector at next time step is calculated using the probability vector at current step and transition probability matrix

$$P_i(n+1) = P_i(n)\pi_i(n). \tag{3.4}$$

Substituting (3.1) and (3.2) in (3.4), the infected probability of node i at the next time step can be derived as

$$P_{i_1}(n+1) = 1 - P_{i_0}(n)(1 - P_{i_{01}}(n)) - P_{i_1}(n)P_{i_{10}}(n). \tag{3.5}$$

We take into account all possible permutations of node i's neighbours' states to derive its transition probabilities as

$$\begin{aligned} P_{i_{01}}(n) = 1 - \sum_{x_{N_i}} &[P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 0) \\ &(1 - \beta_{i,x_{N_i}}(n))], \end{aligned} \tag{3.6}$$

$$\begin{aligned} P_{i_{10}}(n) = \sum_{x_{N_i}} &[P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 1) \\ &\delta(1 - \beta_{i,x_{N_i}}(n))], \end{aligned} \tag{3.7}$$

where $X_{N_i}(n)$ is the state vector of node i's neighbors at time step n; $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 0)$ is the conditional probability given node i is healthy at time step n, its neighbors' state vector at time step n is $x_{N_i}(n)$; $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 1)$ is the conditional probability given node i is infected at time step n, its neighbors' state vector at time step n is $x_{N_i}(n)$; $\beta_{i,x_{N_i}}(n)$ is the conditional probability that given node i's neighbors' states are $x_{N_i}(n)$, it becomes infected at next time step. The formula to calculate this probability was presented before in (3.1) and is equal to

$$\beta_{i,x_{N_i}}(n) = 1 - \prod_{j \in N_i} (1 - \beta_{ji})^{x_j(n)}. \tag{3.8}$$

$P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 0)$ and $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 1)$ characterize the spatial dependence of node i's state to its neighbors' states due to network topology. If these two conditional probabilities are calculated for every node at each time step, we can calculate the exact infected probabilities of nodes at each time step without any approximation. It is remarkable that computing the transition probabilities in (3.6) and (3.7) is computationally too expensive when networks have large numbers of nodes. Therefore, we approximate them to make the SIS model analytically tractable.

3.3 Independent Model

In the independent model, we assume that a node's state is always spatially independent from other nodes' states. Given node i has k neighbors, the total number of its neighbors' states combinations which are required to calculate $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 0)$

and $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 1)$ is reduced from $O(2^k)$ in the full model to $O(k)$ in the independent model. Using the assumption of independence, calculating $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 0)$ and $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 1)$ is simplified and they can be derived as

$$\begin{aligned} & P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 0) \\ & = P(X_{N_i}(n) = x_{N_i}(n)) \\ & = \prod_{j \in N_i} P(X_j(n) = x_j(n)), \end{aligned} \tag{3.9}$$

$$\begin{aligned} & P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 1) \\ & = P(X_{N_i}(n) = x_{N_i}(n)) \\ & = \prod_{j \in N_i} P(X_j(n) = x_j(n)). \end{aligned} \tag{3.10}$$

The assumption of independence causes $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 1)$ and $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 0)$ to become identical in (3.9) and (3.10). Substituting (3.9) and (3.8) into (3.6), we can calculate the transition probability from healthy state to infected state for node i as

$$\begin{aligned} & P_{i_{01}}^{ind}(n) \\ & = 1 - \sum_{x_{N_i}} \prod_{j \in N_i} \left[P(X_j(n) = x_j(n))(1 - \beta_{ji})^{x_j(n)} \right] \\ & = 1 - \prod_{j \in N_i} \sum_{x_j(n)} \left[P(X_j(n) = x_j(n))(1 - \beta_{ji})^{x_j(n)} \right] \\ & = 1 - \prod_{j \in N_i} [1 - \beta_{ji} P_{j_1}(n)]. \end{aligned} \tag{3.11}$$

Also, Substituting (3.10) and (3.8) into (3.7), the transition probability from infected state to healthy state for node i can be written as

$$\begin{aligned} & P_{i_{10}}^{ind}(n) \\ & = \delta \sum_{x_{N_i}} \prod_{j \in N_i} \left[P(X_j(n) = x_j(n))(1 - \beta_{ji})^{x_j(n)} \right] \\ & = \delta \prod_{j \in N_i} \sum_{x_j(n)} \left[P(X_j(n) = x_j(n))(1 - \beta_{ji})^{x_j(n)} \right] \\ & = \delta \prod_{j \in N_i} [1 - \beta_{ji} P_{j_1}(n)]. \end{aligned} \tag{3.12}$$

The derived transition probabilities in (3.11) and (3.12) are the same as analytical results in[3] for the independent model.

3.4 Markov Model

The independent model completely neglects spatial dependence between neighbors' states to simplify the mathematical derivations. In the Markov model, we go one step further compare to the independent model and the spatial dependence of each node's neighbors' states on its state is considered. We assume that neighbors' states of a given node are conditionally independent given its state. Considering the conditionally spatial dependence of neighbors' states, we can write

$$\begin{aligned}&P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 0)\\&= \prod_{j \in N_i} P(X_j(n) = x_j(n)|X_i(n) = 0),\end{aligned} \tag{3.13}$$

$$\begin{aligned}&P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 1)\\&= \prod_{j \in N_i} P(X_j(n) = x_j(n)|X_i(n) = 1).\end{aligned} \tag{3.14}$$

The total number of neighbors' states combinations of a given node i with k neighbors which are required to calculate $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 0)$ and $P(X_{N_i}(n) = x_{N_i}(n)|X_i(n) = 1)$ is reduced from $O(2^k)$ in the full model to $O(k)$ in the Markov model like the independent model. Substituting (3.13) and (3.8) into (3.7), the transition probability from state 0 to state 1 can be written as

$$\begin{aligned}&P_{i01}^{markov}(n) = 1 - \sum_{x_{N_i}} \prod_{j \in N_i}\\&\left[P(X_j(n) = x_j(n)|X_i(n) = 0)(1 - \beta_{ji})^{x_j(n)}\right]\\&= 1 - \prod_{j \in N_i} \sum_{x_j(n)}\\&\left[P(X_j(n) = x_j(n)|X_i(n) = 0)(1 - \beta_{ji})^{x_j(n)}\right]\\&= 1 - \prod_{j \in N_i} [1 - \beta_{ji} P(X_j(n) = 1|X_i(n) = 0)].\end{aligned} \tag{3.15}$$

Substituting (3.14) and (3.8) into (3.7), the transition probability from state 1 to state 0 can be written as

$$\begin{aligned}
&P_{i_{10}}^{markov}(n) = \delta \sum_{x_{N_i}} \prod_{j \in N_i} \\
&\left[P(X_j(n) = x_j(n)|X_i(n) = 1)(1 - \beta_{ji})^{x_j(n)}\right] \\
&= \delta \prod_{j \in N_i} \sum_{x_j(n)} \\
&\left[P(X_j(n) = x_j(n)|X_i(n) = 1)(1 - \beta_{ji})^{x_j(n)}\right] \\
&= \delta \prod_{j \in N_i(n)} [1 - \beta_{ji} P(X_j(n) = 1|X_i(n) = 1)].
\end{aligned} \tag{3.16}$$

Calculating transition probabilities in (3.15) and (3.16) is not as convenient as calculating $P_{i_{01}}^{ind}(n)$ and $P_{i_{10}}^{ind}(n)$. However, in the independent model, computing $P_{i_{01}}^{ind}(n)$ and $P_{i_{10}}^{ind}(n)$ requires only the nodes' infected probabilities, in the Markov model, calculating $P_{i_{01}}^{markov}(n)$ and $P_{i_{10}}^{markov}(n)$ requires calculating $P(X_j(n) = 1|X_i(n) = 0)$ and $P(X_j(n) = 1|X_i(n) = 1)$ in addition to the nodes' infected probabilities. $P(X_j(n) = 1|X_i(n) = 0)$ and $P(X_j(n) = 1|X_i(n) = 1)$ are not the outputs of the epidemic model and we derive a recursive formulas to calculate them. Using two node joint probability $P(X_j(n) = 1, X_i(n) = 1)$, the conditional probabilities of the state of node j given the state of node i can be written as

$$\begin{aligned}
&P(X_j(n) = 1|X_i(n) = 0) \\
&= \frac{P_{j_1}(n) - P(X_j(n) = 1, X_i(n) = 1)}{P_{i_0}(n)},
\end{aligned} \tag{3.17}$$

$$\begin{aligned}
&P(X_j(n) = 1|X_i(n) = 1) \\
&= \frac{P(X_j(n) = 1, X_i(n) = 1)}{P_{i_1}(n)}.
\end{aligned} \tag{3.18}$$

On the other hand, $P(X_j(n+1) = 1, X_i(n+1) = 1)$ can be computed as

$$\begin{aligned}
&P(X_j(n+1) = 1, X_i(n+1) = 1) \\
&= \sum_{u=0}^{1} \sum_{v=0}^{1} P_{uv}^{ji}(n+1) P(X_j(n) = u, X_i(n) = v),
\end{aligned} \tag{3.19}$$

where $P_{uv}^{ji}(n+1) = P(X_j(n+1) = 1, X_i(n+1) = 1|X_j(n) = u, X_i(n) = v)$ is the joint conditional probability of states of nodes i and j to be $(1,1)$ given their states at the previous time index. Consider the case that nodes i and j have the common neighbor node l. In this case, we need to calculate the three joint probability $P(X_j(n) = u, X_i(n) = v, X_l(n))$ to compute $P_{uv}^{ji}(n+1)$. Because this three-node joint probability cannot be calculated using recursive formulas, we neglect it and approximate $P_{uv}^{ji}(n+1)$. First we calculate $P_{00}^{ji}(n+1)$ as

$$P_{00}^{ji}(n+1) \approx I_{ji}(n+1)I_{ij}(n+1), \tag{3.20}$$

where $I_{ji}(n+1) = P(X_j(n+1) = 1|X_j(n) = 0, X_i(n) = 0)$ is the conditional probability that node j is infected in the next time step given nodes i and j are healthy and $I_{ij}(n+1) = P(X_i(n+1) = 1|X_j(n) = 0, X_i(n) = 0)$ is the conditional probability that node i is infected in the next time step given nodes i and j are healthy and. Using β_{ki}'s, $I_{ji}(n+1)$ and $I_{ij}(n+1)$ can be calculated as

$$\begin{aligned}&I_{ji}(n+1) =\\ &1 - \prod_{k \in N_j - \{i\}} [1 - \beta_{kj} P(X_k(n) = 1|X_j(n) = 0)],\end{aligned}$$

$$\begin{aligned}&I_{ij}(n+1) =\\ &1 - \prod_{k \in N_i - \{j\}} [1 - \beta_{ki} P(X_k(n) = 1|X_i(n) = 0)].\end{aligned}$$

Then, we compute $P_{01}^{ji}(n+1)$ as follows:

$$P_{01}^{ji}(n+1) \approx F_{ji}(n+1)E_{ij}(n+1), \tag{3.21}$$

where $F_{ji}(n+1) = P(X_j(n+1) = 1|X_j(n) = 0, X_i(n) = 1)$ is the conditional probability that node j is infected in the next time step given node j is healthy and node i is infected and $E_{ij}(n+1) = P(X_i(n+1) = 1|X_j(n) = 0, X_i(n) = 1)$ is the conditional probability that node i is infected in the next time step given node j is healthy and node i is infected. $F_{ji}(n+1)$ and $E_{ij}(n+1)$ can be calculated as

$$\begin{aligned}&F_{ji}(n+1) =\\ &1 - (1 - \beta_{ij}) \prod_{k \in N_j - \{i\}} [1 - \beta_{kj} P(X_k(n) = 1|X_j(n) = 0)],\end{aligned}$$

$$\begin{aligned}&E_{ij}(n+1) =\\ &1 - \delta \prod_{k \in N_i - \{j\}} [1 - \beta_{ki} P(X_k(n) = 1|X_i(n) = 1)].\end{aligned}$$

Next, we compute $P_{10}^{ji}(n+1)$ as

$$P_{10}^{ji}(n+1) \approx F_{ij}(n+1)E_{ji}(n+1), \tag{3.22}$$

where $F_{ij}(n+1) = P(X_i(n+1) = 1|X_j(n) = 1, X_i(n) = 0)$ is the conditional probability that node i is infected in the next time step given node j is infected and node i is healthy and $E_{ji}(n+1) = P(X_j(n+1) = 1|X_j(n) = 1, X_i(n) = 0)$ is the conditional probability that node j is infected in the next time step given nodes j is infected and node i is healthy. $F_{ij}(n+1)$ and $E_{ji}(n+1)$ can be calculated as follows:

$$F_{ij}(n+1) = \\ 1 - (1-\beta_{ji}) \prod_{k \in N_i - \{j\}} [1 - \beta_{ki} P(X_k(n) = 1|X_i(n) = 0)],$$

$$E_{ji}(n+1) = \\ 1 - \delta \prod_{k \in N_j - \{i\}} [1 - \beta_{kj} P(X_k(n) = 1|X_j(n) = 1)].$$

Finally, $P_{11}^{ji}(n+1)$ is calculated as

$$P_{11}^{ji}(n+1) \approx R_{ji}(n+1) R_{ij}(n+1), \tag{3.23}$$

where $R_{ji}(n+1) = P(X_j(n+1) = 1|X_j(n) = 1, X_i(n) = 1)$ is the conditional probability that node j is infected in the next time step given nodes i and j are infected and $R_{ij}(n+1) = P(X_i(n+1) = 1|X_j(n) = 1, X_i(n) = 1)$ is the conditional probability that node i is infected in the next time step given nodes i and j are infected. $R_{ji}(n+1)$ and $R_{ij}(n+1)$ can be calculated as

$$R_{ji}(n+1) = \\ 1 - \delta(1-\beta_{ij}) \prod_{k \in N_j - \{i\}} [1 - \beta_{kj} P(X_k(n) = 1|X_j(n) = 1)],$$

$$R_{ij}(n+1) = \\ 1 - \delta(1-\beta_{ji}) \prod_{k \in N_i - \{j\}} [1 - \beta_{ki} P(X_k(n) = 1|X_i(n) = 1)].$$

Substituting (3.20), (3.21), (3.22), and (3.23) into (3.19) yields a recursive formula to calculate $P(X_j(n) = 1|X_i(n) = 0)$ and $P(X_j(n) = 1|X_i(n) = 1)$.

3.5 SIMULATION RESULTS

In this section, we compare the Markov and the independent model with Monte Carlo simulation results for various topologies and different epidemic parameters.

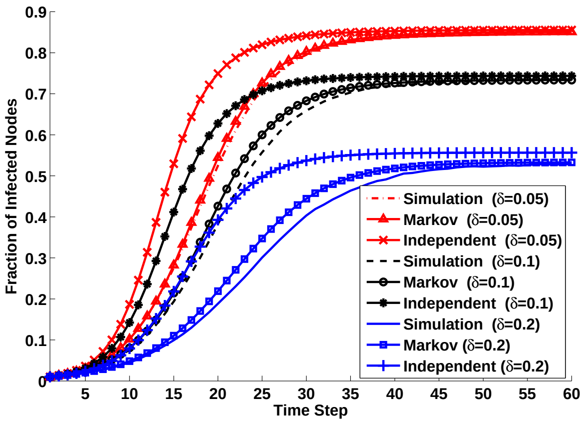

Figure 3.1: *Malware propagation in ER random graph with 1000 nodes,* $< k >= 4$*, and* $\beta = 0.1$.

Fig. 3.1 compares predictions of the Markov model and the independent model with simulation results for the ER network with 1000 nodes and average node degree of 4 ($< k >= 4$). In all cases, $\beta = 0.1$ and malware epidemic starts with an initial .01 percent of nodes infected ($i_0 = .01$). The number of iterations for each case is 100 times. In all three cases ($\delta = 0.2$,$\delta = 0.1$,$\delta = 0.06$), the Markov model yields the closer results to simulation results than the independent model's results. Also, it can be seen that an increase in epidemic strength ($\frac{\beta}{\delta}$) causes the independent model's curve to become closer to the curves of the Markov model and simulations. It means that, as much as the epidemic strength decreases, the gain of using the Markov model instead of the independent model becomes more substantial.

In Fig. 3.2, we show that the fraction of initial infected nodes has an important role on the accuracy of the independent and the Markov models' predictions. In all cases, malware propagation runs on the ER random graph with $N = 1000$, $< k >= 8$, $\beta = .1$, and $\delta = .34$. The number of iteration for each case is 100. When the initial fraction of infected nodes decreases, the difference between performance of the Markov model and the independent model increases and the advantage of using the Markov model becomes more significant. In all four cases ($i_0 = .001$, $i_0 = .01$, $i_0 = .1$, $i_0 = .4$), the Markov model results in more accurate results than the ones from the independent model.

The level of connectivity in the networks beside the epidemic strength determines how fast an epidemic spreads.[17] shows that the connectivity level in a given network can be characterized by its adjacency matrix's largest eigenvalue. They prove that the epidemic spreads out if the epidemic strength is greater than the inverse of largest eigenvalue of adjacency matrix λ_{max} otherwise epidemic dies out. The epidemic strength and λ_{max} determine the speed of malware propagation in a given network[?] . In Fig. 3.3, we study the effect of the connectivity level of ER network with 1000 nodes on the performances of the independent and the Markov models. Keeping $\frac{\beta}{\delta}\lambda_{max}$ as a constant number causes this comparison to become fair. In all three cases ($< k >= 4$, $< k >= 8$, $< k >= 16$), $\beta = .1$ and we set the

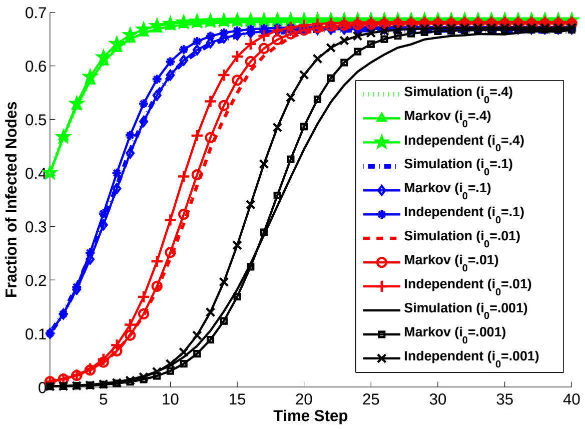

Figure 3.2: *Malware propagation in ER random graph with 1000 nodes,* $< k >= 8$, $\beta = 0.1$, $\delta = .34$.

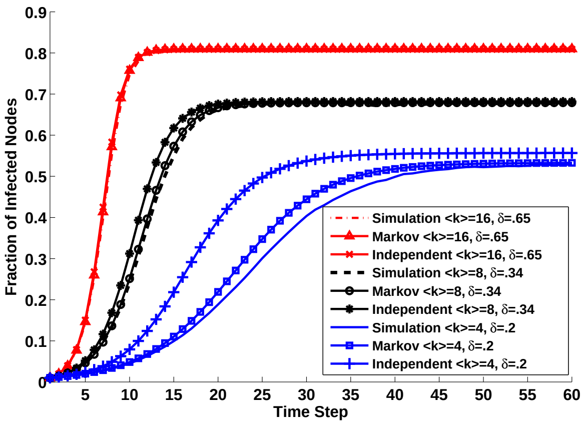

Figure 3.3: *Malware propagation in ER random graph with 1000 nodes,* $\beta = .1$, *and* $\frac{\beta}{\delta}\lambda_{max} = 2.6457$.

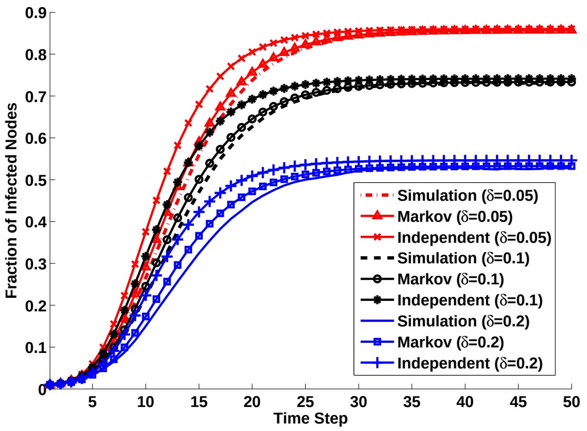

Figure 3.4: *Malware propagation in BA network with 1000 nodes, $\beta = .1$, $< k >= 4$, and $i_0 = .01$.*

value of δ in order to have $\frac{\beta}{\delta}\lambda_{max} = 2.6457$. Increasing the average node degree causes the independent model's results become closer to the Markov model's results and simulation results, while $\frac{\beta}{\delta}\lambda_{max}$ is same for all the three cases. It can be seen that a decrease of the average node degree, which is equivalent to decrease of λ_{max}, degrades the performance of the independent model, whereas the Markov model still predicts the epidemic evolution with better accuracy.

Fig. 3.4 compares the simulation results with the results of the independent model and Markov model in the Barabasi-Albert (BA) network with 1000 nodes, $< k >= 4$, $\beta = .1$, and $i_0 = .01$. The simulation curves are resultants of 100 times iterations. In all four cases ($\delta = .2$, $\delta = .1$, $\delta = .05$), the Markov model outperforms the independent model.

In Fig. 3.5, we show that when average node degree of a BA network with 1000 nodes increases, predictions of the Markov and the independent models become more accurate. In all three cases ($< k >= 2$, $< k >= 4$, $< k >= 6$), $\beta = .1$, $\delta = .1$, $i_0 = .01$, Markov model's results match the simulation results better than the independent model's results. Employing the Markov model has more significant gain for less connected BA networks, similarly to the ER networks cases.

Fig. 3.6 depicts epidemic evolution with $i_0 = .01$ and $\beta = .14$ on the Oregon graph. Oregon graph is a real topology which is collected from the Oregon router views[1] and can be used to simulate propagation of malware that targets at attacking routers in the Internet. This network has 11461 nodes and 32370 links. As you can see in both cases ($\delta = .24, \delta = .08$), the two models yields result pretty close to the simulation results. However, the Markov model's predictions are slightly more accurate.

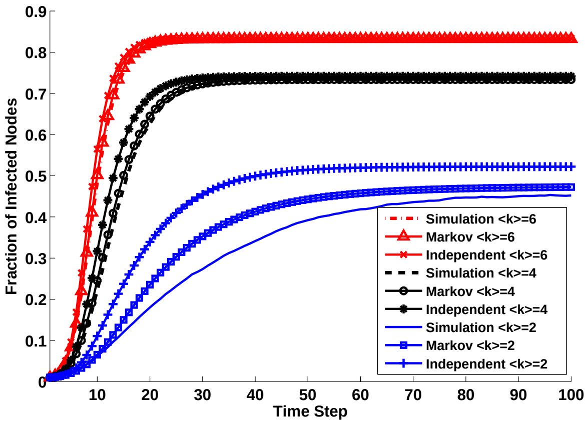

Figure 3.5: *Malware propagation in BA network with 1000 nodes, $\beta = .1$, $\delta = .1$, and $i_0 = .01$.*

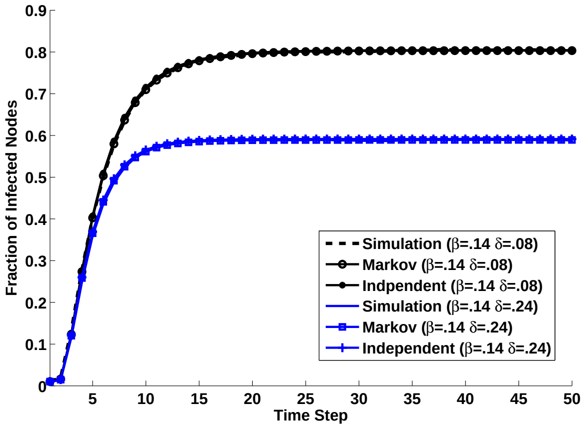

Figure 3.6: *Malware propagation in Oregon graph with $i_0 = .01$ and $\beta = .14$.*

3.6 CONCLUSION

In this chapter, we present a SIS malware propagation modeling in discrete time that can characterize the spreading of random and topological scanning malware. This model assumes the dependence of the cure probability to nodes' states. First, we develop the full model which takes into account the exact spatial dependence between neighbors' states. When networks have large numbers of nodes, computing the probabilities that characterizes the full spatial dependence between nodes becomes computationally too expensive. Therefore, we derive the independent model which is based on the assumption of spatial independence of neighbors' states. Finally, we propose the Markov model which considers partial spatial dependence between neighbors' states. We show that the Markov model yields far better performance than the independent model in the transient behavior of malware spreading. We show that small initial fraction of infected nodes, low connectivity level of topologies, and low epidemic strength makes the gain of exploiting of Markov models instead of independent models more substantial.

Chapter 4

Epidemic Threshold in Dynamic Networks

4.1 Introduction

Epidemics typically start with some initial infected nodes. Infected nodes can cause their healthy neighbors to become infected with some probability. With time and in some cases with external intervention, infected nodes can be cured and go back to a healthy state. The study of epidemic dispersals on networks aims at explaining how epidemics evolve and spread in networks. One of the most interesting questions regarding an epidemic spread in a network is whether the epidemic dies out or results in a massive outbreak. Epidemic threshold is a parameter that addresses this question by considering both the network topology and epidemic strength. Although epidemic spread in static networks has been studied extensively in recent years[3,4,8–10], dispersal of epidemics and epidemic threshold derivation in dynamic networks has received little attention[11,13,15,16]. It is remarkable that in most realistic cases of epidemic spread, the underlying networks are dynamic and links between nodes are functions of time. For example in the area of human disease epidemics, contact networks between people are not fixed and change over time because people continuously move from one location to another location. Another example is malware propagation in mobile adhoc networks. The movement of users results in dynamic topologies. Many Bluetooth devices are becoming susceptible to viruses such as Cabir and Comm Warrior. Another classic example of dynamic networks is underlying networks for the spread of diseases between animals or plants where the factors that influence the spread of disease-carrying spores are typically dynamic.

In this chapter, we consider the SIS (Susceptible-Infected-Susceptible) model for epidemic spread. In this model, healthy nodes can become infected through their infected neighbors and infected nodes have the probability to become cured. We should notice that in the SIS model, an infected node's state after being cured is susceptible again. We assume that infected nodes have the same cure probability of δ and every infected node can make its healthy neighbor infected with infectious probability β. We study switching dynamic networks where at each step the adjacency matrices of networks are randomly chosen from sets of matrices. First, the nonlinear dynamical system of infection probabilities of nodes based on the assumption of independence among the state of nodes is developed. Then, we prove that the origin is always one equilibrium point of this time-varying dynamical system and its stability depends on the network topology and the values of δ and β. After that, the linearized version of the nonlinear epidemic system is derived to determine whether the origin is asymptotically stable or not. We show that if the origin is not a stable equilibrium of the system, the epidemic spreads out, otherwise it dies out. Then, the joint spectral radius of a set of matrices is defined. In Theorem 1, we employ the concept of the joint spectral radius to derive the analytical epidemic threshold for dynamic networks. In Theorem 2, the simplified version of epidemic threshold for undirected networks is derived. Since the epidemic threshold for undirected networks depends only on the largest spectral radius of a set of system matrices, evaluation of the epidemic threshold is computationally less expensive compare to directed networks. In corollary 1, it is shown that the derived epidemic threshold confirms the conventional analytical results for static networks. Then, the proposed epidemic threshold for dynamic networks is extended to the case of periodic networks. Moreover, we study epidemic spread in dynamic regular networks and show that the epidemic threshold for dynamic regular networks is the same as static regular networks. An upper bound for the probability of an epidemic spreading out in dynamic Gilbert networks is derived. Finally, we simulate epidemics in Watts-Strogatz, Barabasi-Albert, Regular, and dynamic Gilbert networks to validate our analytical results. Additionally, we examine our theoretical results in the context of real networks by considering the MIT reality mining graphs.

The rest of the chapter is organized as follows. In section 4.2, we review related prior works on epidemic threshold in dynamic networks. Section 4.3 contains the analytical results for epidemic threshold in general for dynamic networks as well as the simplified epidemic thresholds for special cases of dynamic networks. In Section 4.4, we use simulation results to validate our theoretical analysis.

4.2 Related Work

In[16], the epidemic threshold in the case of a SIR (Susceptible-Infected-Recovered) model is derived for a simple class of dynamic random networks. In this class of dynamic networks, the number of neighbors of a given node is fixed, but its neighbors change stochastically as a Poisson process through instantaneous neighbor exchanges. Continually, pairs of edges are chosen randomly with equal probability and instantaneously interchanged. In[13], the authors present a model describing a SIS epidemic on dynamic networks using set of ordinary differential equations. The SIS effective degree model for a static contact network in[9] is modified by introducing link activation and deletion rates. They calculate the epidemic threshold for this model and show that limiting the maximum nodal degree of a network can be used to prevent the outbreak of an epidemic. In[11], the authors derive the epidemic threshold for dynamic networks with alternating (periodic) adjacency matrices. They consider the SIS model for epidemic propagation in networks and show that if the dynamic behavior of a time-varying network can be characterized by T repeating alternating graphs, $L = \{A_1, A_2, ..., A_T\}$. The system matrix, S, of this dynamical system can be expressed as

$$S = \prod_{i=1}^{T} [(1-\delta)I + \beta A_i)], \tag{4.1}$$

where A_i's dimension is $n \times n$ (n is the number of nodes), I is an n by n identity matrix, and δ and β denote, respectively, cure probability and infectious probability. They prove that if the spectral radius of the system matrix is less than one, the origin is an asymptotically stable equilibrium point of the system and the epidemic dies out. It is noticeable that this result holds for the cases when there are repeating patterns of adjacency matrices and the order of repetition is preserved. In[15], the same authors as[11] study malware propagation on mobile ad hoc networks. They extend their result for epidemic threshold of periodic networks in[11] to general cases in which repeating order of adjacency matrices can be arbitrary. In Theorem I (Mobility model threshold) of their chapter, they state that if a mobility model can be represented as a sequence of connectivity graphs $L = \{A_1, A_2, ..., A_T\}$, one adjacency matrix A_t for each index $t \in \{1, 2, .., T\}$, then the epidemic threshold is

$$\tau = \lambda_S, \tag{4.2}$$

where λ_S is the largest eigenvalue of the matrix S defined in (4.1).

This Theorem claims that the condition for asymptotic stability of the origin for a given dynamic network whose adjacency matrices at each index can be arbitrarily chosen from a set of matrices is that the spectral radius of the matrix S is less than one, which is different from our analytical results in Theorem 1, to be presented later, for the same assumptions.

4.3 ANALYTICAL RESULTS

In this section, we develop a dynamical system for epidemic spread based on the assumption of spatial independence between states of nodes in a given network. Then, a linearized version of the dynamical system is derived to determine the epidemic threshold. Next, employing the joint spectral radius, we quantify epidemic threshold for dynamic networks. Moreover, it is proved that the epidemic threshold in undirected networks only depends on the maximum spectral radius of the set of system matrices. Then, we extend the results of proposed epidemic threshold to static and periodic networks. In addition, we show that the epidemic threshold for dynamic regular networks is the same as static regular networks. Finally, we calculate an upper bound for the probability of epidemic spreading in dynamic Gilbert networks.

Using the assumption of spatial independence between states of nodes in a given network, we can write the infection probability of each node in the network as

$$p_i(t+1) = 1 - p_i(t)\delta - (1 - p_i(t)) \prod_{j \in N_i(t)} [1 - p_j(t)\beta], \tag{4.3}$$

where $N_i(t)$ denotes the set of neighbors of node i at index t, which is a function of time. The infection probabilities of nodes can be interpreted as state variables of a dynamical system. (4.3) shows infection probabilities at a given index are nonlinear functions of infection probabilities of the previous index. Therefore, the epidemic dynamical system is nonlinear. The corresponding state space of this nonlinear system is subspace $[0,1]^n$ in R^n, where n is the number of nodes in the network. For instance, when $n = 2$, the state space is a rectangle whose vertices are points of $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ in R^2. Given initial infection probabilities of nodes, we can calculate the evolving trajectory of infection probabilities in state space. The family of evolving trajectories of states in the state space is called a phase portrait. Studying the steady state behavior of dynamical systems requires finding the equilibrium points. If P^* is an equilibrium point, $P^*(t+1) = P^*(t) = P^*$. So, we can write

$$p_i^* = 1 - p_i^*\delta - (1 - p_i^*) \prod_{j \in N_i(t)} \left[1 - p_j^*\beta\right], \tag{4.4}$$

where p_i^* is the infection probability of the i^{th} node when the system state reaches the equilibrium point. (4.4) is the equilibrium equation corresponding to node i. To find the equilibrium points of a given epidemic system with n nodes, we need to solve a system of n equations with n unknowns. In the case of dynamic networks, this system of equations is changing with time and equilibrium points by definition are static points which satisfy this system of equations for all time. An epidemic dynamical system may have more than one equilibrium point. It is obvious from (4.4) that the origin is the trivial solution. It means that for all different values of β and δ and for any arbitrary topologies of networks, the origin is always an equilibrium point. However, the values of β and δ with topologies of networks determine the stability status of the origin. If the origin is an asymptotically stable equilibrium point, the epidemic dies out. On the other hand, the epidemic spreads out when the origin is an unstable equilibrium point. For the case when the epidemic spreads out, there is another equilibrium point in addition to the origin. In this case, this non-origin equilibrium point is asymptotically stable and by finding it, we can calculate the final fraction of infected nodes.

4.3.1 Linearization of System Equations

One way to identify the stability status of an equilibrium point of a nonlinear system is to study the stability of the linearized system at that equilibrium point. In the case of epidemic networks, we are interested in determining the stability status of the origin because if the origin is an asymptotically stable point, the epidemic dies out only if no other equilibrium points exist in the subspace $[0,1]^n$, otherwise asymptotic stability is only local. Therefore,

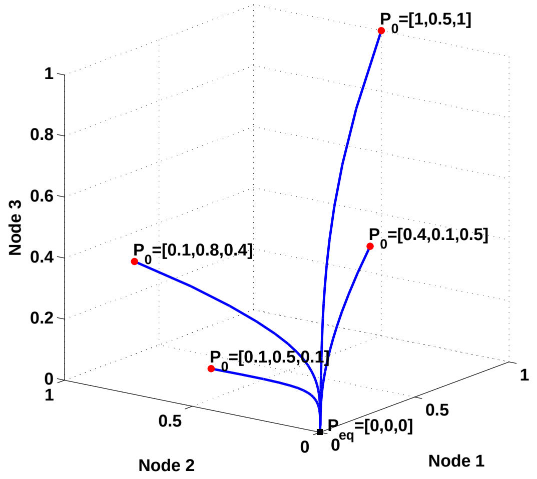

Figure 4.1: *Phase portrait of an epidemic networks with three nodes while epidemic dies out.*

we linearize the epidemic nonlinear system at the origin. Neglecting nonlinear terms in (4.3), we can write

$$p_i(t+1) = p_i(t)(1-\delta) + \sum_{j \in N_i(t)} p_j(t)\beta. \tag{4.5}$$

We can rewrite (4.5) in format of a matrix equation as

$$P(t+1) = [(1-\delta)I + \beta A_t]\, P(t), \tag{4.6}$$

where $P(t) = [p_1(t)p_2(t)...p_n(t)]^T$ is the system state or, in the other words, the vector of infection probabilities of nodes at index t, A_t is the adjacency matrix at index t, and I denotes an $n \times n$ identity matrix. From this point, we call $M_t = [(1-\delta)I + \beta A_t]$ the system matrix at index t.

4.3.2 Epidemic Threshold

“Under what conditions do epidemics die out?,” is the most important question that can be asked in the study of epidemics. As we mentioned earlier, we need to study the stability of the origin in the epidemic dynamical system to answer this question. If the origin is an asymptotically stable equilibrium point, the system state reaches the origin and infection probabilities of all nodes become zero and remain zero. Fig. 4.1 depicts trajectories of state evolution of an epidemic network for different initial infection probabilities. The network has three nodes and its adjacency matrix is static. Node 1 is connected to node 2 and 3, but there is no link between node 2 and node 3. Each axis of this 3D plot represents the infection probability of one of the three nodes over time. The dots with the P_0 label in

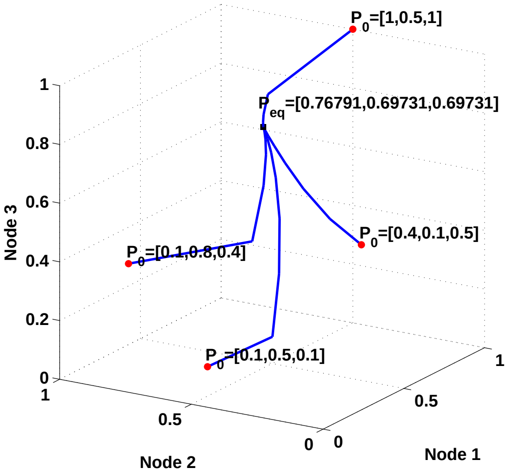

Figure 4.2: *Phase portrait of an epidemic networks with three nodes while epidemic spreads out.*

state space represent the initial infection probabilities and the dot with the P_{eq} represents the equilibrium point of the system. $\delta = .2$ and $\beta = .1$. As you can see in this figure, all trajectories reach the origin regardless of their initial states. It shows that in this case the origin is asymptotically stable and epidemics die out. In the case of spreading epidemics, the origin is not a stable equilibrium point and the state variables converge to a non-zero equilibrium point and remain in that point. This equilibrium point determines the final fraction of infected nodes. Fig. 4.2 shows the trajectories of state evolution of the same network as Fig. 4.1. The initial states are the same as the ones in Fig. 4.1 and the only difference is the value of β. In this case, $\beta = .6$. As you can see in this figure, the epidemic spreads out and for all different infection probability vectors, the system state reaches the equilibrium point $P_{eq} = [0.76791, 0.69731, 0.69731]$ whose elements represent, respectively, the infection probability of nodes 1, 2, and 3 in steady state. Before tackling the problem of stability of the origin, we need to present some definitions.

Definition 1. *Given M is a set of matrices, define*

$$\widehat{\rho}_k(M, ||.||) := sup \left\{ \left|\left| \prod_{i=1}^{k} M_i \right|\right| : M_i \in M \ for \ 1 \leq i \leq k \right\},$$

where $\widehat{\rho}_k(M)$ is the largest possible norm of all products of k matrices chosen in the set M. The joint spectral radius $\widehat{\rho}(M)$ is defined as[12]

$$\widehat{\rho}(M) := \lim_{k\to\infty} \widehat{\rho}_k(M, ||.||)^{\frac{1}{k}}. \tag{4.7}$$

Therefore, the joint spectral radius of set M is the maximum possible norm of products of matrices in the set M when number of products k goes to infinity.

Definition 2. *Given M is a set of matrices, define*

$$\overline{\rho}_k(M) := sup\left\{\rho\left(\prod_{i=1}^{k} M_i\right) : M_i \in M \text{ for } 1 \leq i \leq k\right\},$$

where ρ denoted the spectral radius and $\overline{\rho}_k(M)$ is the largest possible spectral radius of all products of k matrices chosen in the set M. The generalized spectral radius $\overline{\rho}(M)$ is defined as[5]

$$\overline{\rho}(M) := \lim_{k\to\infty} \overline{\rho}_k(M)^{\frac{1}{k}}. \tag{4.8}$$

In[2], the authors proved that for a bounded set of matrices, the generalized spectral radius is equal to the joint spectral radius.

Lemma 1. ***Four members inequality***[6] *For a given arbitrary set of matrices M and any $k \geq 1$*

$$\overline{\rho}_k(M)^{\frac{1}{k}} \leq \overline{\rho}(M) \leq \widehat{\rho}(M) \leq \widehat{\rho}_k(M)^{\frac{1}{k}},$$

independent of the induced norm used to define $\widehat{\rho}_k(M)$.

Let us consider a set L of all possible adjacency matrices A_i and at each time instant the adjacency matrix is randomly chosen from this set. L is surely bounded and may be finite or infinite. We define M as the set of system matrices corresponding to the adjacency matrices in L. M_i is a member of the set M and defined as $M_i = [(1-\delta)I + \beta A_i]$. Therefore, M is also bounded. If L is finite, M is finite too and if M is infinite, L is infinite.

Theorem 1. *Consider a set L of all possible adjacency matrices of a dynamic network, infectious probability of β, and cure probability of δ. If the joint spectral radius of set M of system matrices is less than one, the origin is an asymptotically stable equilibrium point and the epidemic dies out.*

Proof. If $\widehat{\rho}(M) < 1$, we can write, by definition 1,

$$\lim_{k\to\infty} \widehat{\rho}_k(M)^{\frac{1}{k}} < 1.$$

If we raise the both sides of the above inequality to the power k, we can conclude that

$$\lim_{k\to\infty} \widehat{\rho}_k(M) = 0.$$

Considering the formula of $\widehat{\rho}_k(M)$ in Definition. 1, for any product of matrices $M_i \in M$, we can write

$$0 \leq \left\|\prod_{i=1}^{k} M_i\right\| \leq \widehat{\rho}_k(M).$$

We showed that the right hand side of the above inequality goes to zero when $k \to \infty$. Therefore for any product of M_is, we can write

$$\lim_{k\to\infty} \left|\left|\prod_{i=1}^{k} M_i\right|\right| = 0.$$

Considering this fact that $||A|| = 0 \Leftrightarrow A = 0$, if $\lim_{k\to\infty} \left|\left|\prod_{i=1}^{k} M_i\right|\right| = 0$, we can write

$$\lim_{k\to\infty} \prod_{i=1}^{k} M_i = 0.$$

If $\lim_{k\to\infty} \prod_{i=1}^{k} M_i = 0$, for any initial infection probability vector $P(0)$ we can write

$$\lim_{k\to\infty} \left[\prod_{i=1}^{k} M_i\right] P(0) = \lim_{k\to\infty} \left[\prod_{i=1}^{k} [(1-\delta)I + \beta A_i]\right] P(0) = 0,$$

which shows that the origin is an asymptotically stable equilibrium point for any initial infection probability vector and any random sequence of adjacency matrices if the joint spectral radius is less than one. In this case, the final infection probability vector is zero and epidemic dies out.

□

Theorem 2. *Consider a set L of all possible adjacency matrices of a dynamic network with undirected graphs, set M of the system matrices corresponding to set L of the adjacency matrices, infectious probability of β, and cure probability of δ. If the largest spectral radius of the matrices in set M is less than one, the origin is asymptotically stable equilibrium point and the epidemic dies out.*

Proof. If the network graph is undirected, its corresponding adjacency matrix is symmetric so that its corresponding system matrix is also symmetric. We know that for a given symmetric matrices M_i, we can calculate the induced 2 norm of M_i as follows

$$||M_i||_2 = \sqrt{\rho(M_i^T M_i)} = \sqrt{\rho(M_i)^2} = \rho(M_i).$$

If we use the induced 2 norm to calculate $\widehat{\rho}_1(M)$ and $\overline{\rho}_1(M)$, respectively we can write

$$\widehat{\rho}_1(M) = sup\left\{||M_i||_2 : M_i \in M\right\} = sup\left\{\rho(M_i) : M_i \in M\right\}$$

and

$$\overline{\rho}_1(M) = sup\left\{\rho(M_i) : M_i \in M\right\}.$$

Therfore, we can conclude that for a set of symmetric matrices

$$\widehat{\rho}_1(M) = \overline{\rho}_1(M) = sup\left\{\rho(M_i) : M_i \in M\right\}. \tag{4.9}$$

Moreover, we mentioned in Lemma 1 that the four member inequality holds for any $k \geq 1$. Therefore, we can write

$$\overline{\rho}_1(M) \leq \overline{\rho}(M) \leq \widehat{\rho}(M) \leq \widehat{\rho}_1(M). \tag{4.10}$$

Considering (4.9) and (4.10), we can write

$$\overline{\rho}(M) = \widehat{\rho}(M) = sup\{\rho(M_i) : M_i \in M\}.$$

Therefore the joint spectral radius of set M of symmetric matrices is equal to the largest spectral radius of matrices in the set. Based on Theorem 1, we can conclude that if the largest spectral radius of the system matrices of an undirected dynamic network is less than one, the origin is asymptotically stable equilibrium point and the epidemic dies out. □

Corollary 1. *Consider a static epidemic network with adjacency matrix A, infectious probability of β, and cure probability of δ. The epidemic dies out if $\frac{\beta}{\delta} < \frac{1}{\rho(A)}$.*

Proof. For a static network, M, the set of system matrices, has only one element which is $(1-\delta)I + \beta A$. In this case, $\overline{\rho}_k(M)$, the largest possible spectral radius of all products of k matrices chosen in the set M, can be written as

$$\overline{\rho}_k(M) = sup\left\{\rho\left(\prod_{i=1}^{k} M_i\right) : M_i \in M\right\} = \rho((1-\delta)I + \beta A)^k.$$

M is a bounded set and it is proved in[2] that for a bounded set of matrices the joint spectral radius is equal to the generalized spectral radius. Hence, we can calculate the joint spectral radius as

$$\widehat{\rho}(M) = \lim_{k\to\infty} \overline{\rho}_k(M)^{\frac{1}{k}} = \rho((1-\delta)I + \beta A).$$

Based on Theorem 1, epidemic dies out if the joint spectral radius of the set of system matrices is less than one. For a static network, the joint spectral radius is equal to $1 - \delta + \beta\rho(A)$. Therefore, epidemic dies out if

$$\frac{\beta}{\delta} < \frac{1}{\rho(A)}. \tag{4.11}$$

□

The epidemic threshold for static networks in (4.11) is the same as analytical results for static networks' epidemic threshold in[3].

Corollary 2. *Consider a dynamic network with a fixed repetition pattern of T adjacency matrices in a set $L = \{A_1, A_2, ..., A_T\}$, infectious probability of β, and cure probability of δ. The epidemic dies out if*

$$\rho(\prod_{i=1}^{T}[(1-\delta)I+\beta A_i]) < 1.$$

Proof. Consider a dynamic network with a fixed repetition pattern of T adjacency matrices and $k = mT$ where m is a positive integer. For the case where $k = mT$, $\widehat{\rho}_k(M)$ can be written

$$\begin{aligned}\overline{\rho}_k(M) &= sup\left\{\rho\left(\prod_{i=1}^{mT} M_i\right)\right\}\\ &= sup\left\{\rho\left(\prod_{i=1}^{m}\left[\prod_{i=1}^{T}[(1-\delta)I+\beta A_i]\right]\right)\right\}\\ &= \rho\left(\prod_{i=1}^{T}[(1-\delta)I+\beta A_i]\right)^m.\end{aligned} \tag{4.12}$$

Since the set of system matrices is bounded, its joint spectral radius is equal to its generalized spectral radius. Hence, we can calculate the joint spectral radius as

$$\widehat{\rho}(M) = \lim_{m\to\infty}\overline{\rho}_{mT}(M)^{\frac{1}{mT}} = \rho\left(\prod_{i=1}^{T}[(1-\delta)I+\beta A_i]\right)^{\frac{1}{T}}.$$

According to Theorem 1, if the joint spectral radius of the set of system matrices is less than one, the epidemic dies out. Therefore, in this case epidemic dies out if $\rho\left(\prod_{i=1}^{T}[(1-\delta)I+\beta A_i]\right)^{\frac{1}{T}} < 1$ or equivalently

$$\rho(\prod_{i=1}^{T}[(1-\delta)I+\beta A_i]) < 1. \tag{4.13}$$

□

The derived epidemic threshold for the periodic dynamic networks is the same as the one in[11].

In the next corollary, we propose the condition under which epidemics die out in the case of dynamic networks with all elements of set L correspond to regular networks. In regular dynamic networks, although the links between nodes are dynamic and a given node can change its neighbors at each index, all nodes have the same node degree and they preserve their node degrees.

Corollary 3. *The epidemic in a dynamic regular network with undirected graphs and node degree of $\overline{k}$ dies out if $\frac{\beta}{\delta} < \frac{1}{\overline{k}}$.*

Proof. We know that the spectral radius of a regular symmetric graph is equal to its node degree. Considering $M_i = (1-\delta)I + \beta A_i$, we can calculate the spectral radius of M_i for any regular adjacency matrix as follows:

$$\rho(M_i) = 1 - \delta + \beta\rho(A_i) = 1 - \delta + \beta\overline{k}.$$

Therefore, all system matrices have the same spectral radius of $1 - \delta + \beta\overline{k}$ and the largest spectral radius of the system matrices is also equal to $1 - \delta + \beta\overline{k}$. Based on Theorem 2, we can conclude that epidemics in dynamic networks with regular undirected graphs die out if the largest spectral radius of the system matrices is less than one. Hence, the epidemic dies out if $1 - \delta + \beta\overline{k} < 1$ or equivalently if

$$\frac{\beta}{\delta} < \frac{1}{\overline{k}}. \tag{4.14}$$

□

It is remarkable that (4.14) is the same as the epidemic threshold for static regular networks[3].

Up to this point, we studied the dynamic networks whose adjacency matrices are deterministic and given. However, for the cases that adjacency matrices are stochastic and not deterministic, the joint spectral radius of the set of the system matrices is a random variable. Therefore, the condition for the dying out of the epidemics turns out to be in terms of statistical characteristics of the joint spectral radius. For instance, adjacency matrix's elements of a given dynamic Gilbert network have the same Bernoulli distribution. In these types of networks, at each index links between pairs of nodes exists with the probability of P. In the next corollary, we derive the upper bound for the probability of spreading out of epidemics for Gilbert dynamic networks in terms of expected value of the joint spectral radius. To prove corollary 4, first we need to state theorem 3.

Theorem 3. *Consider $\widehat{M} = \prod_{i=1}^{k} M_i$ where M_i denotes the system matrix corresponding to one realization of a Gilbert dynamic network's adjacency matrix. For the matrix $\widehat{M}$, the expected value of summation of each column's elements is*

$$E\left\{\sum_{q=1}^{N} |\widehat{m}_{q,n}|\right\} = [1 - \delta + (N-1)\beta P]^{k}\,\forall n = 1, 2, ..., N, \tag{4.15}$$

where P is the probability of link existence and $\widehat{m}_{q,n}$ denotes the element in q^{th} row and n^{th} column of matrix $\widehat{M}$.

Proof. We prove this Theorem via induction. In the case of a Gilbert dynamic network, off-diagonal elements of adjacency matrix A are independent and identically distributed (iid) Bernoulli random variables with parameter P. The first step is to show that (4.15) is correct when $k = 1$. Assume $k = 1$. In this case, $\widehat{M} = (1-\delta)I + \beta A$ and $E\left\{\sum_{q=1}^{N} |\widehat{m}_{q,n}|\right\}$ can be written as

$$E\left\{\sum_{q=1}^{N} |\widehat{m}_{q,n}|\right\} = (1-\delta) + \beta \sum_{i=1}^{N-1} E\{X_i\}, \tag{4.16}$$

where the X_i's are iid random variables with parameter P. $E\{X_i\} = P$. Therefore, we can rewrite (4.16) as

$$E\left\{\sum_{q=1}^{N} |\widehat{m}_{q,n}|\right\} = (1-\delta) + (N-1)\beta P. \tag{4.17}$$

(4.17) shows that (4.15) is correct for $k = 1$. The second step is to assume (4.15) is correct for k and prove it for $k+1$. Assume $\widehat{M} = \prod_{i=1}^{k} M_i$. Considering the assumption of correctness of (4.15) for k, $E\left\{\sum_{q=1}^{N} |\widehat{m}_{q,n}|\right\} = [1-\delta+(N-1)\beta P]^k$. Suppose $R = M_{k+1}\widehat{M}$. $r_{q,n}$, the element in the the q^{th} row and the n^{th} column of R can be written in terms of the elements of $\widehat{M}$ as

$$r_{q,n} = (1-\delta)\widehat{m}_{q,n} + \beta \sum_{j=1, j\neq q}^{N} X_j \widehat{m}_{j,n}, \tag{4.18}$$

where the X_j's are iid Bernoulli random variables with parameter P. Therefore, we can write $\sum_{q=1}^{N} |r_{q,n}|$ as

$$\sum_{q=1}^{N} |r_{q,n}| = (1-\delta)\sum_{q=1}^{N} \widehat{m}_{q,n} + \beta \sum_{q=1}^{N}\left[\sum_{j=1, j\neq q}^{N} X_j \widehat{m}_{j,n}\right], \tag{4.19}$$

where X_j's and $\widehat{m}_{j,n}$ are independent. Hence, $E\left\{\sum_{q=1}^{N} |r_{q,n}|\right\}$ can be written as

$$\begin{aligned} E\left\{\sum_{q=1}^{N} |r_{q,n}|\right\} &= (1-\delta)E\left\{\sum_{q=1}^{N} \widehat{m}_{q,n}\right\} + \\ & P\beta E\left\{\sum_{q=1}^{N}\left[\sum_{j=1, j\neq q}^{N} \widehat{m}_{j,n}\right]\right\}. \end{aligned} \tag{4.20}$$

On the other hand, $E\left\{\sum_{q=1}^{N}\left[\sum_{j=1, j\neq q}^{N} \widehat{m}_{j,n}\right]\right\} = (N-1)E\left\{\sum_{q=1}^{N} \widehat{m}_{q,n}\right\}$. Considering $E\left\{\sum_{q=1}^{N} \widehat{m}_{q,n}\right\} = [(1-\delta)+\beta P(N-1)]^k$, we can rewrite (4.20) as

$$\begin{aligned} E\left\{\sum_{q=1}^{N} |r_{q,n}|\right\} &= [(1-\delta)+\beta P(N-1)]\, E\left\{\sum_{q=1}^{N} \widehat{m}_{q,n}\right\} \\ &= [(1-\delta)+\beta P(N-1)]^{k+1}. \end{aligned} \tag{4.21}$$

The result in (4.21) for $k+1$ is the last step in the proof of this Theorem through induction. □

Corollary 4. *For a given dynamic Gilbert network with N nodes and the probability of existence of links of P, the probability that epidemic spreads out is upper bounded by $[1-\delta+(N-1)\beta P]$.*

Proof. We define $\widehat{M}$ as $\widehat{M}=\prod_{i=1}^{k} M_i$ where M_i denotes the system matrix corresponding to one realization of a Gilbert dynamic network's adjacency matrix. $\widehat{m}_{q,n}$ denotes the element in q^{th} row and n^{th} column of matrix $\widehat{M}$. In the Appendix, we show that for all columns of $\widehat{M}$

$$E\left\{\sum_{q=1}^{N}|\widehat{m}_{q,n}|\right\}=[1-\delta+(N-1)\beta P]^{k}\,\forall n=1,2,...,N, \tag{4.22}$$

where E denotes the expected value. Considering $\left|\left|\widehat{M}\right|\right|_1=\max_n\sum_{q=1}^{N}|\widehat{m}_{q,n}|$ and (4.22), we can calculate $E\left\{\left|\left|\widehat{M}\right|\right|_1\right\}$ as

$$E\left\{\left|\left|\widehat{M}\right|\right|_1\right\}=[1-\delta+(N-1)\beta P]^{k}\,.$$

Because the above equality holds for any product of M_is, $E\left\{\widehat{\rho}_k(M)\right\}$ can be written as

$$E\left\{\widehat{\rho}_k(M)\right\}=[1-\delta+(N-1)\beta P]^{k}\,.$$

And consequently, the expected value of the joint spectral radius can be calculated as

$$E\left\{\widehat{\rho}(M)\right\}=E\left\{\lim_{k\to\infty}\widehat{\rho}_k(M)^{\frac{1}{k}}\right\}=[1-\delta+(N-1)\beta P]\,.$$

We employ the Markov inequality and the expected value of the joint spectral radius to compute the upper bound for the probability of the joint spectral radius to be more than one. Using the Markov inequality, we can write

$$Prob(\widehat{\rho}(M)\geq 1)\leq E\{\widehat{\rho}(M)\}. \tag{4.23}$$

Substituting the expected value of the joint spectral radius in (4.23), we can conclude

$$Prob(\widehat{\rho}(M)\geq 1)\leq[1-\delta+(N-1)\beta P]\,. \tag{4.24}$$

Based on Theorem 1, the epidemic dies out if the joint spectral radius is less than one. Therefore, the probability of epidemic spreads out is equal to the probability of the joint spectral radius to be greater than one. Considering (4.24), we can conclude that the probability of epidemic to spread out is upper bounded by $[1-\delta+(N-1)\beta P]$. □

We should notice that when $[1-\delta+(N-1)\beta P]$ is greater than one, it is not informative. Therefore, we consider the $min\{1,[1-\delta+(N-1)\beta P]\}$ as the upper bound for the probability of spreading out.

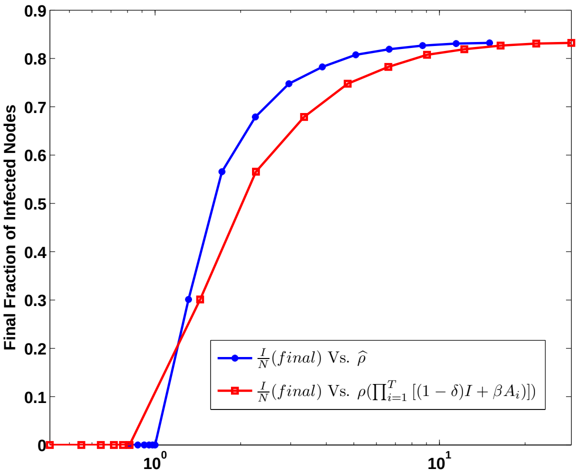

Figure 4.3: *Final fraction of infected nodes for dynamic Watts-Strogatz networks with* 1000 *nodes and rewiring probability of* .5.

4.4 SIMULATION RESULTS

In this section, we validate our theoretical results via simulation of epidemic on synthetic and real dynamic networks. First, we simulate an epidemic on a dynamic Watts-Strogatz network and compare the derived epidemic threshold with the threshold proposed in[15]. Then, the simulation result of final fraction of infected nodes versus the joint spectral radius for a dynamic Barabasi-Albert network is presented. Moreover, we evaluate our analytical results in the case of real networks by simulating an epidemic on the set of extracted graphs from MIT Reality Mining data set. Next, epidemic on a dynamic regular network is simulated. Finally, we validate the derived upper bound for the probability of epidemic spreading in a dynamic Gilbert network using simulation results.

In Fig. 4.3, $\frac{I}{N}$ denotes the final fraction of infected nodes where I denotes the final number of infected nodes and N denotes the total number of nodes. Fig. 4.3 plots $\frac{I}{N}$ versus the joint spectral radius of the system matrices and the spectral radius of the products of the system matrices for a dynamic Watts-Strogatz network with 1000 nodes and rewiring probability of .5. To realize a dynamic Watts-Strogatz, the adjacency matrix of network at each index is chosen randomly from a set of four Watts-Strogatz graphs with average node degrees of 4, 8, 12, and 16 and, respectively, spectral radii of 4.46242, 8.41081, 12.40911, and 16.38739. The adjacency matrix with average node degree of 16 has the largest spectral radius among the matrices in the set. Therefore, the joint spectral radius of the set of system matrices is equal to the spectral radius of the system matrix corresponding to the adjacency matrix with average node degree of 16. Although the set of adjacency matrices is fixed during simulation, we change the value of $\frac{\beta}{\delta}$ to generate different cases of epidemic strength. We keep the value of $\delta = 0.2$ as a constant while increasing β from 0.00052 to 0.86652. The

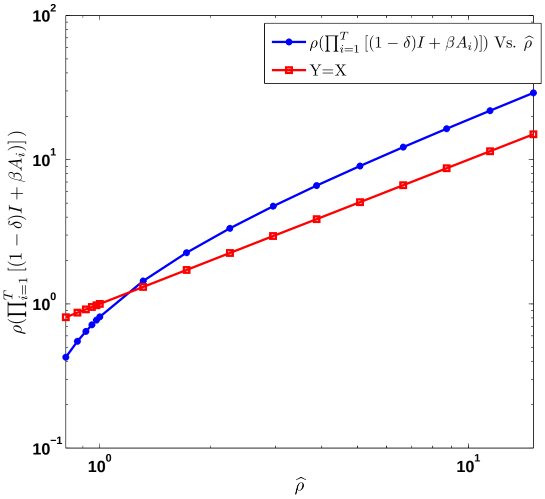

Figure 4.4: *Comparison between the joint spectral radius and the spectral radius of the product of the system matrices for dynamic Watts-Strogatz networks in Fig. 4.3.*

number of iteration for each case is 20. As it can be observed from the curve of the final fraction of infected nodes versus the joint spectral radius, epidemics die out for all the cases that the value of the joint spectral radius is less than one. As soon as the value of the joint spectral radius increases beyond one, epidemics spread out, which confirms the analytical results of Theorems 1 and 2. The curve of the final fraction of infected nodes versus the spectral radius of the system matrices product shows that $\rho(\prod_{i=1}^{T}[(1-\delta)I+\beta A_i)])$ is not the accurate epidemic threshold. It can be observed that epidemic spreads out for some values of the spectral radius of the product which are less than one. This simulation results contradicts the analytical results of Theorem 1 in[15] which says that if the spectral radius of the product is less than one, the epidemic dies out. Fig. 4.4 compares the spectral radius of the product of system matrices of the dynamic Watts-Strogatz network with the joint spectral radius of its set of system matrices. In this figure, the curve of $Y = X$ with unity slope helps us to determine for which values of the joint spectral radius, the spectral radius of the product is greater than the joint spectral radius or vice versa. It can be seen in this figure that up to the specific value of the joint spectral radius, the spectral radius of the product is less than the joint spectral radius. In other words, there are some points that the joint spectral radius is greater than one and the spectral radius of the product is less than one. This results in wrong predictions of dying out epidemics for those cases if we choose the spectral radius of the product as the epidemic threshold.

Fig. 4.5 shows the fraction of infected nodes over time for a dynamic Barabasi-Albert network with 1000 nodes. To realize a dynamic Barabasi-Albert network, we select four Barabasi-Albert graphs with average node degree of 4, 8, 12, and 16 with, respectively, spectral radii of 12.66217, 17.36462, 22.36887, and 27.91071 as the set of adjacency matrices. During simulation, the adjacency matrix of this dynamic network at each index is randomly chosen from this set of matrices with equal probability. We change the value of β while

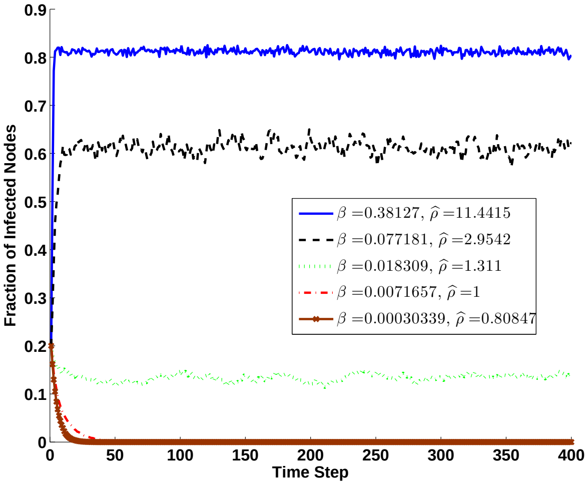

Figure 4.5: *Fraction of infected nodes over time for dynamic Barabasi-Albert network with* 1000 *nodes.*

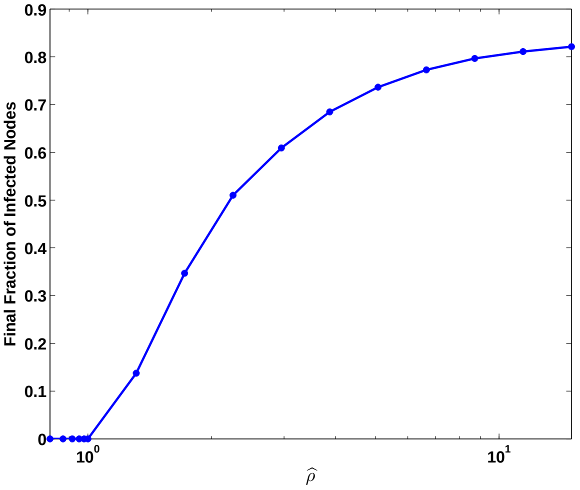

Figure 4.6: *Final fraction of infected nodes for dynamic Barabasi-Albert network with* 1000 *nodes Vs. the joint spectral radius.*

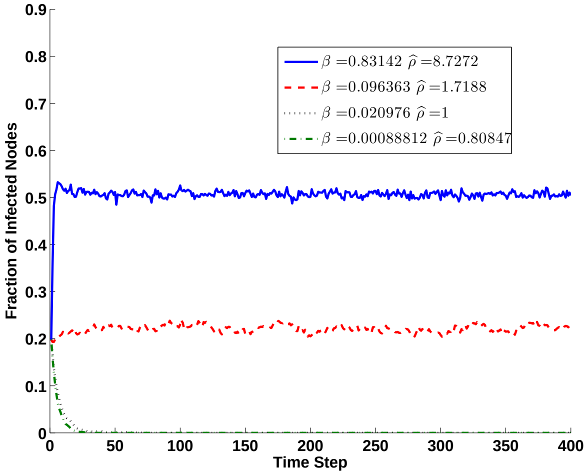

Figure 4.7: *Fraction number of infected nodes for MIT Reality Mining dynamic over time.*

fixing the value of $\delta = 0.2$. The number of iteration for all cases is 20 and the initial fraction of infected nodes is 0.2. According to Theorem 2, the joint spectral radius of undirected dynamic networks is equal to $[1 - \delta + \beta\rho(A_j)]$, where A_j refers to the adjacency matrix with the largest value of spectral radius in the set of adjacency matrices. In the case of this dynamic Barabasi-Albert network, the largest spectral radius is 27.91071. As you can see in this figure, increasing the value of β leads to an increase in the joint spectral radius and, eventually, the final fraction of infected nodes. Also, epidemics die out for the case the joint spectral radius is less than one. Fig. 4.6 depicts the final fraction of infected nodes for the mentioned dynamic Barabasi-Albert network. Here $\delta = 0.2$ for all cases while β increases from 0.00038 to 0.63481. It can be seen in this figure that epidemics spread out for the cases where the joint spectral radius is greater than one.

Fig. 4.7 shows the fraction of infected nodes over time for a dynamic network whose time-varying adjacency matrices are extracted from the MIT Reality Mining data set[7]. This data set contains the adjacency connectivity matrix of 94 persons using mobile phones pre-installed with different special software including the logger of Bluetooth devices which was triggered when two mobile phones' distance was approximately five meters or less. Bluetooth scans were carried out every 5 minutes. This data set contains the collected information from mobile phones from September 2004 to June 2005. We extract eight adjacency matrices for 8 consecutive hours from 9 o'clock in the morning to 4 o'clock in the afternoon of September 1, 2004. These eight adjacency matrices have, respectively, the spectral radii of 6.30117, 5.41546, 9.44439, 9.09696, 8.36535, 9.53451, 9.05251, and 7.41181. We simulated epidemic dispersal over the dynamic network whose adjacency matrix was randomly chosen from the eight adjacency matrices at each index. For all four cases, $\delta = 0.2$ and the initial fraction of infected nodes was 0.2. The number of iterations for each case is 50. As you can see in this figure, increasing β causes an increase in the value of the joint spectral radius and epidemic spreads out when the joint spectral radius is greater than one. Fig. 4.8 depicts the final

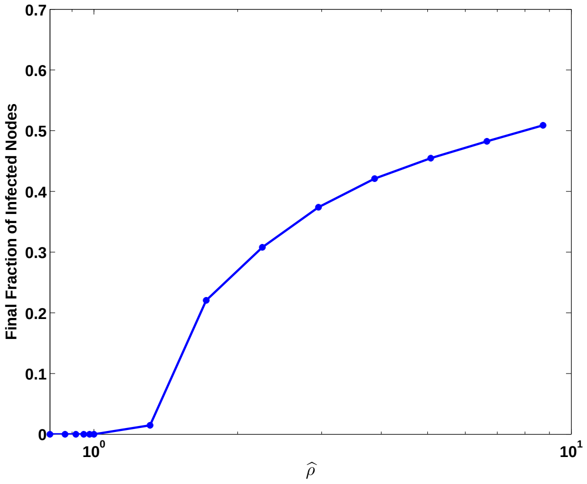

Figure 4.8: *Final fraction of infected nodes for MIT Reality Mining dynamic Vs. the joint spectral radius.*

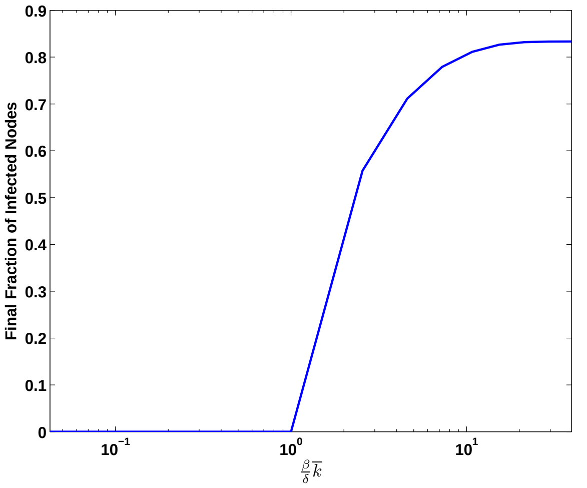

Figure 4.9: *final fraction of infected nodes for a dynamic regular network with* 1000 *nodes and node degree of* 8*.*

fraction of infected nodes for the dynamic network in Fig. 4.7. We fix the value of $\delta = 0.2$ while increasing the value of β from 0.00089 to 0.83142. The number of iteration for each case is equal to 50. As it can be observed in this figure, the epidemic spreads out when the joint spectral radius is greater than one. For the cases that the epidemic spreads out, the final fraction of infected nodes increases with the increase in the value of joint spectral radius.

Fig. 4.9 shows the final fraction of infected nodes for a dynamic regular network with 1000 nodes and node degree of $\overline{k} = 8$ versus the product of $\frac{\beta}{\delta}$ and node degree. In this network, although every node's neighbor set is a function of time and nodes' neighbors

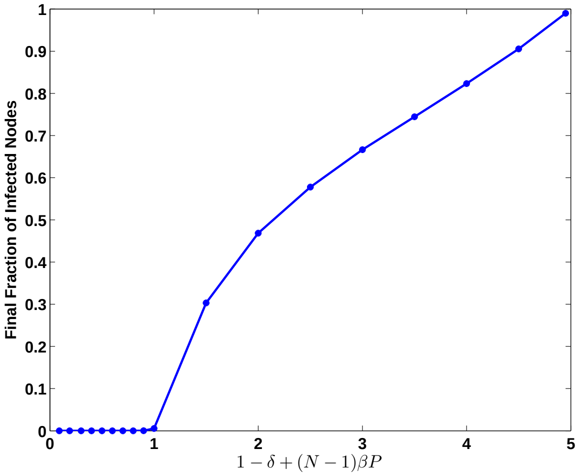

Figure 4.10: *final fraction of infected nodes for a dynamic Gilbert network with* 1000 *nodes and node degree of* 8*.*

change at each index, the nodes' degree is fixed and the same over time. During the simulation of this dynamic regular network, the node degree is set to 8. However, the nodes' neighbor sets are dynamic. We simulate epidemics for different values of β in the interval of $[0.00106, 0.99089]$ while fixing the value of $\delta = 0.2$. The number of iteration for each case is 100. As you can see in Fig. 4.9 epidemics die out when $\frac{\beta}{\delta} < \frac{1}{k}$, which confirms the result of corollary 3.

Fig. 4.10 depicts the final fraction of infected nodes for a dynamic Gilbert network with 1000 nodes and a probability of connection of $P = 0.004$ versus the values of $[1 - \delta + (N-1)\beta P]$ which is the upper bound for the probability of an epidemic spreading out in a Gilbert dynamic networks. To realize a dynamic Gilbert network, we generate a new Gilbert adjacency matrix with 1000 nodes and connection probability of 0.004 at each step of simulation so that the links between nodes turn to be dynamic and functions of time. During simulation, values of δ and β are chosen based on the Table 4.1. Also, Table 4.1 shows the value of $[1 - \delta + (N-1)\beta P]$ for the corresponding values of δ and β. As it can be seen in Fig. 4.10, the epidemic dies out up to the point where the upper bound of the probability of the epidemic spreading out is less than one. When this upper bound reaches one, the epidemic starts to spread out, which confirms the result of corollary 4. We should notice that although having an upper bound greater than one for a probability is not informative, the upper bound of the probability of spreading out epidemics in dynamic Gilbert networks can be used as a measure of the epidemic strength. This is evident in Fig. 4.10, as increasing value of upper bound leads to an increase in the final fraction of infected nodes.

Table 4.1: *δ and β values used in simulation of the epidemic in the dynamic Gilbert network*

δ	β	$1-\delta+(N-1)\beta P$
0.95	0.01	0.09
0.85	0.01	0.19
0.74	0.01	0.3
0.64	0.01	0.4
0.54	0.01	0.5
0.44	0.01	0.6
0.34	0.01	0.7
0.24	0.01	0.8
0.14	0.01	0.9
0.04	0.01	1
0.7	0.3	1.5
0.6	0.4	2
0.5	0.5	2.5
0.4	0.6	3
0.3	0.7	3.5
0.2	0.8	4
0.1	0.9	4.5
0.01	0.99	4.95

4.5 CONCLUSION

In this chapter, SIS epidemics spread in dynamic networks is studied. We propose an analytical way to derive the analytical epidemic threshold, which can be applied to any dynamic network whose adjacency matrix is randomly chosen from a set of matrices at any index whereas the previous results calculate epidemic threshold only for periodic dynamic networks whose adjacency matrices have fixed repetition patterns. The linearized version of the nonlinear epidemic system to derive the epidemic threshold is employed. We show that an epidemic dies out if the origin is an asymptotically stable equilibrium point. We derive the epidemic threshold for dynamic networks using the joint spectral radius of the system matrices. We calculate the simplified version of epidemic threshold for undirected dynamic networks using the fact that the joint spectral radius of a set of symmetric matrices is equal to the largest spectral radius of matrices in that set. Then, the epidemic threshold for dynamic regular networks is derived. For dynamic Gilbert networks, we compute the upper bound of the probability of an epidemic spreading out in terms of the expected value of the joint spectral radius. Finally, we show the accuracy of our analytical results using simulation results of epidemics in Watts-Strogatz, Barabasi-Albert, Regular, MIT Reality Mining and Gilbert dynamic networks.

Bibliography

[1] University of oregon route views project [online]. available: http://routeviews.org/.

[2] M. A. Berger and Y. Wang. Bounded semigroups of matrices. *Linear Algebra and its Applications*, 166:21–27, 1992.

[3] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.*, 10, 2008.

[4] Z. Chen and C. Ji. Spatial-temporal modeling of malware propagation in networks. *IEEE Transactions on Neural Networks*, 16:1291–1303, 2005.

[5] I. Daubechies and J. C. Lagarias. Sets of matrices all infinite products of which converge. *Linear Algebra Appl.*, 161:227–263, 1992.

[6] I. Daubechies and J. C. Lagarias. Corrigendum/addendum to: Sets of matrices all infinite products of which converge. *Linear Algebra Appl.*, 327:69–83, 2001.

[7] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *Proc. of the National Academy of Sciences*, 2009.

[8] A. Ganesh, L. Massoulie, and D. Towsley. The effect of network topology on the spread of epidemics. *INFOCOM, 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, 2:1455– 1466, 2005.

[9] J. Lindquist, J. Ma, P. V. D. Driessche, and F. H. Willeboordse. Effective degree network disease models. *Journal of Mathematical Biology*, 62:143–164, 2010.

[10] P. V. Mieghem, J. Omic, and R. Kooij. Virus spread in networks. *IEEE/ACM Transactions on Networking*, 17:1–14, 2009.

[11] B. A. Prakash, H. Tong, N. Valler, M. Faloutsos, and C. Faloutsos. Virus propagation on time-varying networks: theory and immunization algorithms. *In Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases*, 2010.

[12] G. C. Rota and G. Strang. A note on the joint spectral radius. *Nederl. Akad. Wet., Proc., Ser. A*, 63:379–381, 1960.

[13] M. Taylor, T. J. Taylor1, and I. Z. Kiss. Epidemic threshold and control in a dynamic network. *Physical Review E 85*, 2012.

[14] K. Thomas and D. M. Nicol. The koobface botnet and the rise of social malware. *5th International Conference on Malicious and Unwanted Software (MALWARE)*, pages 63–70, 2010.

[15] N. C. Valler, B. A. Prakash, H. Tong, M. Faloutsos, and C. Faloutsos. Epidemic spread in mobile ad hoc networks: Determining the tipping point. *NETWORKING 2011, Lecture Notes in Computer Science, Springer Berlin*, 6640:266–280, 2011.

[16] E. Volz and L. A. Meyers. Epidemic thresholds in dynamic contact networks. *Journal of the Royal Society Inteface 6*, 2009.

[17] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: an eigenvalue viewpoint. *in Proceedings of 22nd International Symposium on Reliable Distributed Systems*, pages 25–34, 2003.

Appendix A

Error Prediction

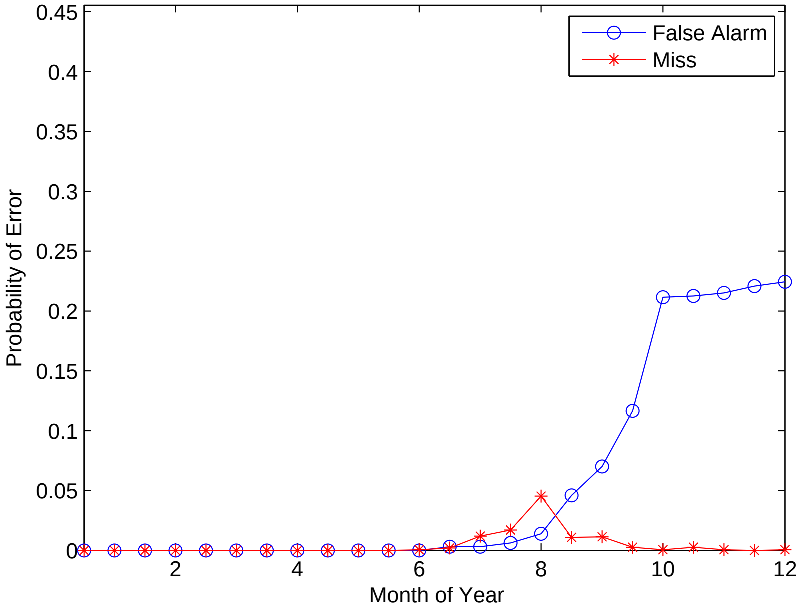

Figure A.1: *Errors prediction for 2005 based on observations from 2006.*

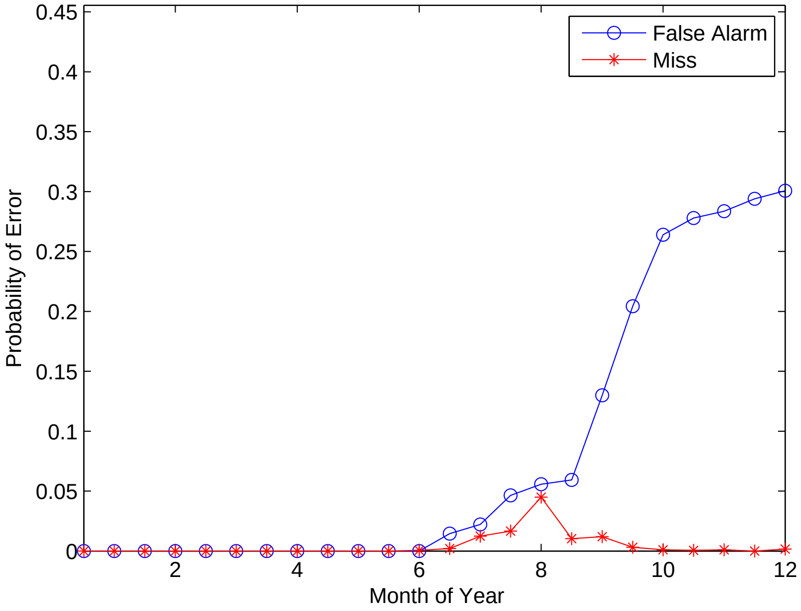

Figure A.2: *Errors prediction for 2005 based on observations from 2007.*

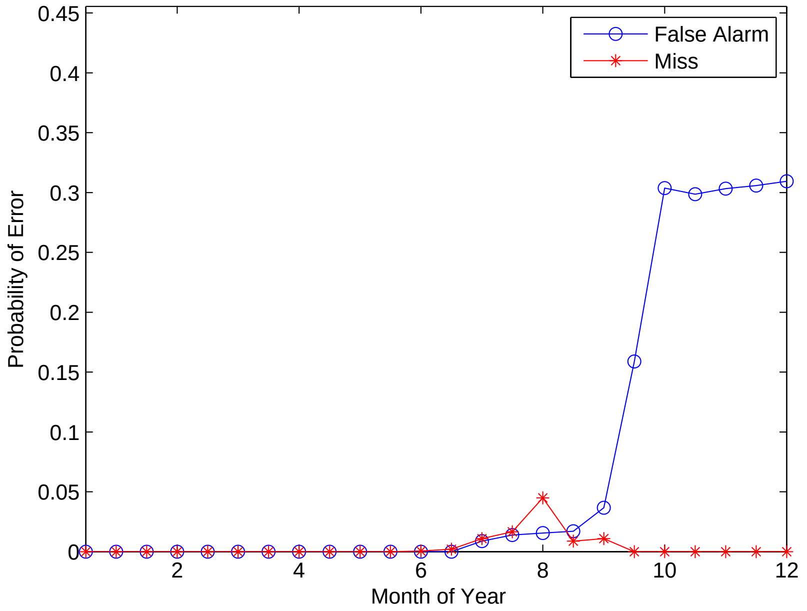

Figure A.3: *Errors prediction for 2005 based on observations from 2008.*

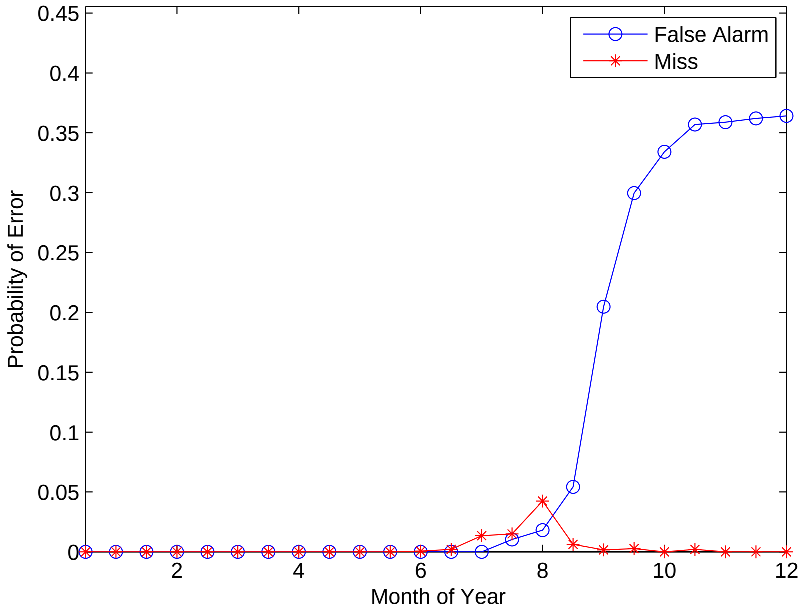

Figure A.4: *Errors prediction for 2005 based on observations from 2009.*

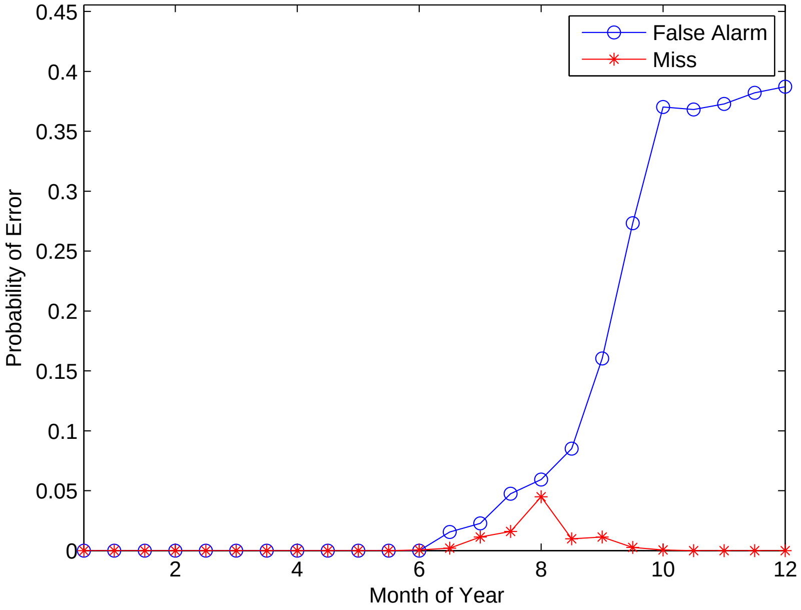

Figure A.5: *Errors prediction for 2005 based on observations from 2006 and 2007.*

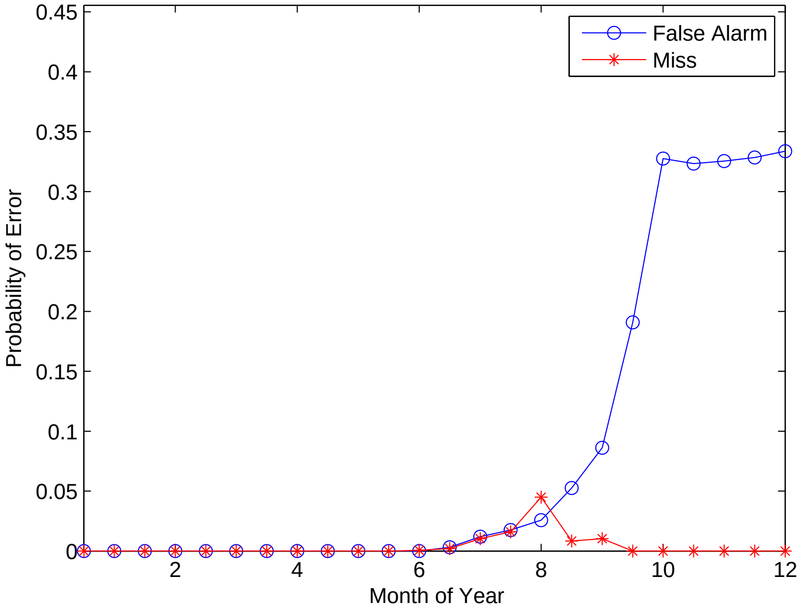

Figure A.6: *Errors prediction for 2005 based on observations from 2006 and 2008.*

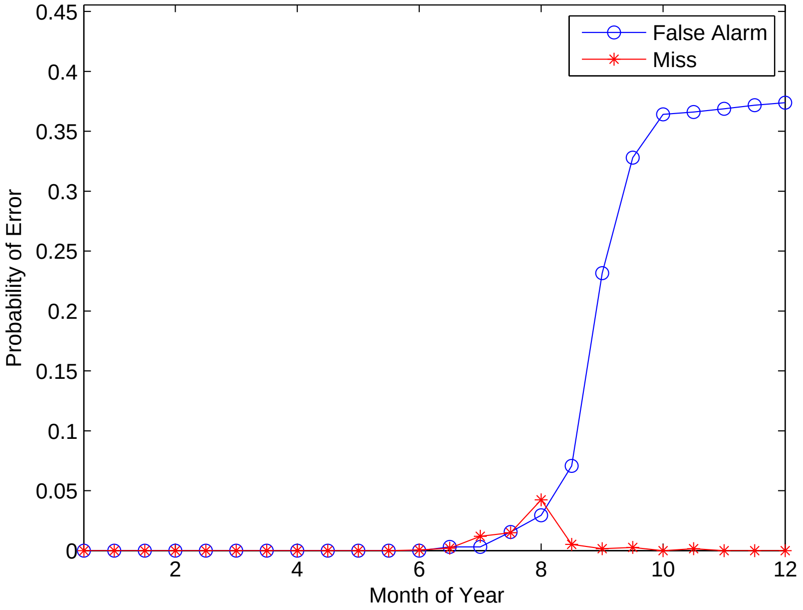

Figure A.7: *Errors prediction for 2005 based on observations from 2006 and 2009.*

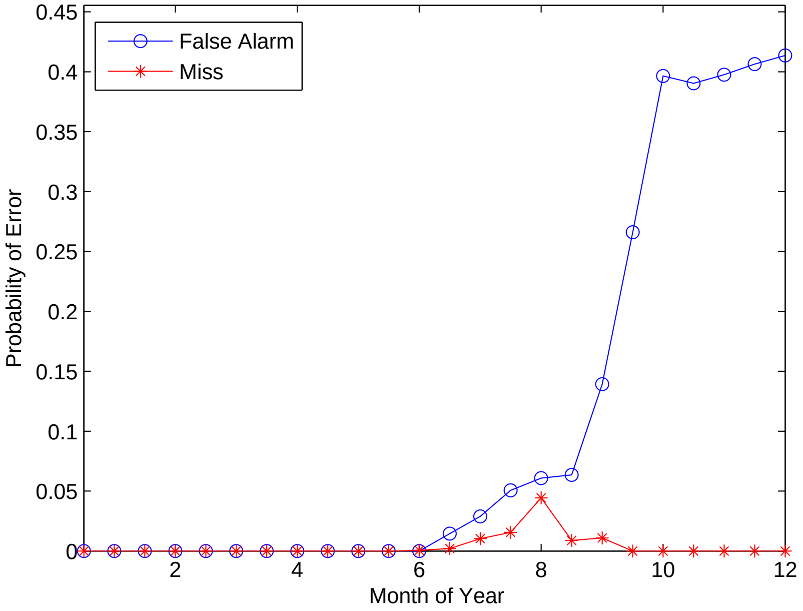

Figure A.8: *Errors prediction for 2005 based on observations from 2007 and 2008.*

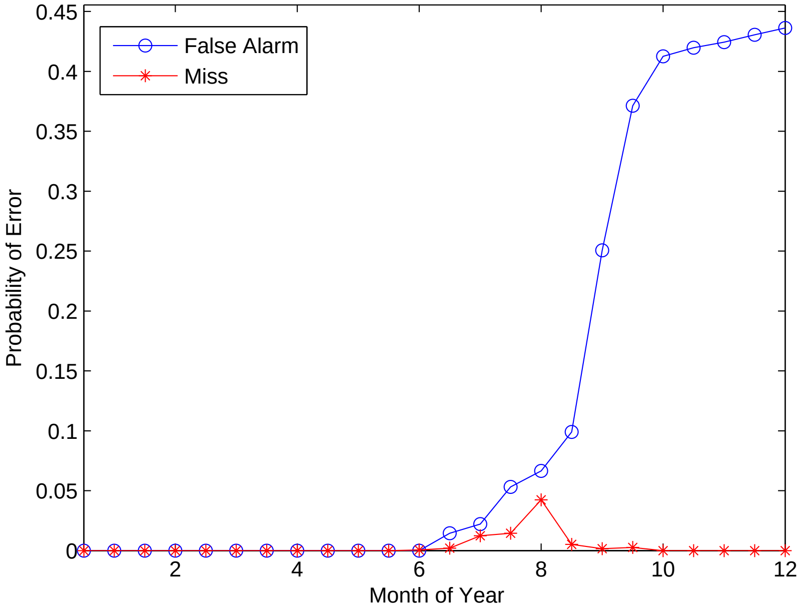

Figure A.9: *Errors prediction for 2005 based on observations from 2007 and 2009.*

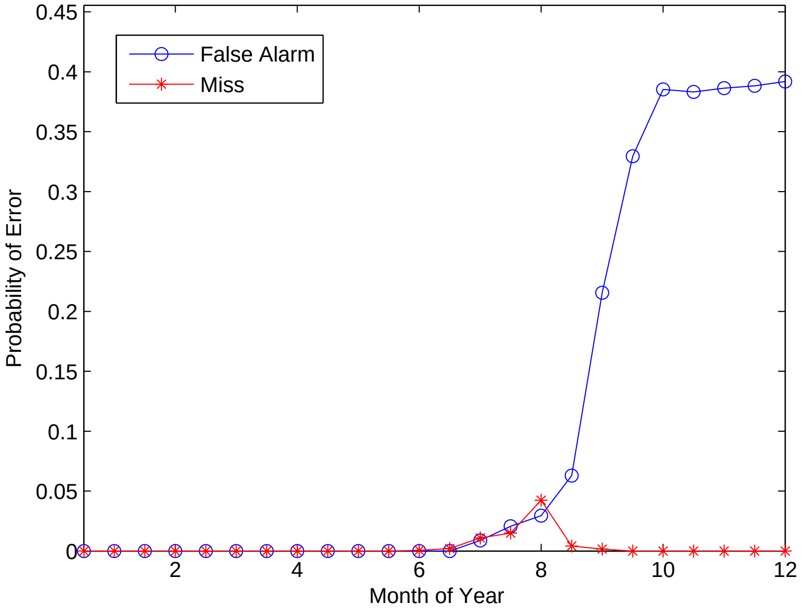

Figure A.10: *Errors prediction for 2005 based on observations from 2008 and 2009.*

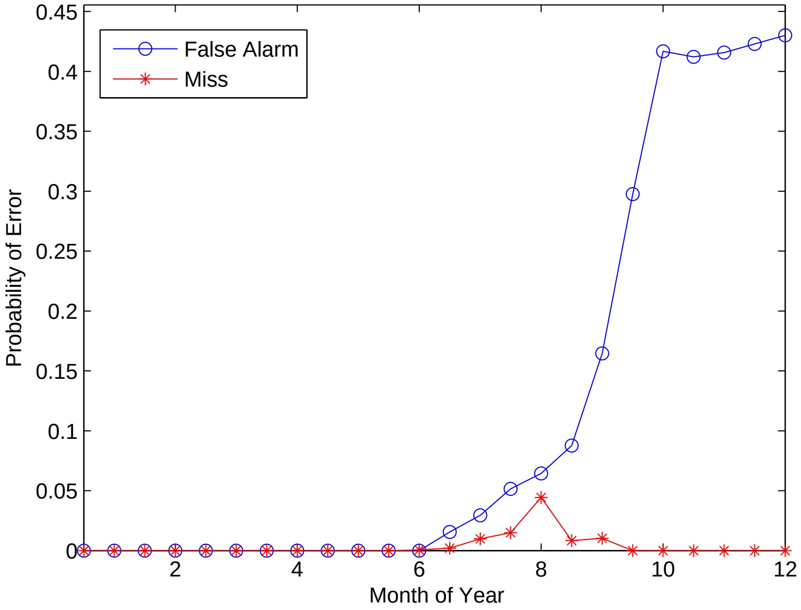

Figure A.11: *Errors prediction for 2005 based on observations from 2006, 2007, and 2008.*

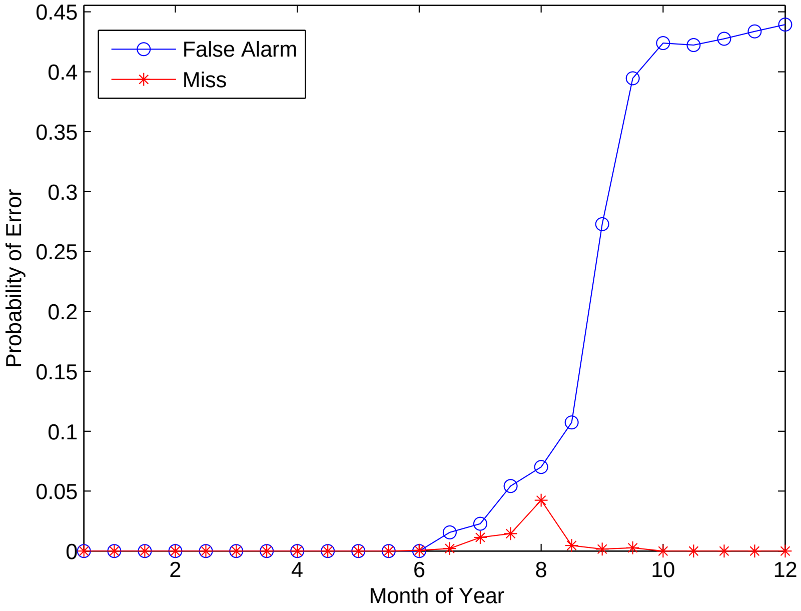

Figure A.12: *Errors prediction for 2005 based on observations from 2006, 2007, and 2009.*

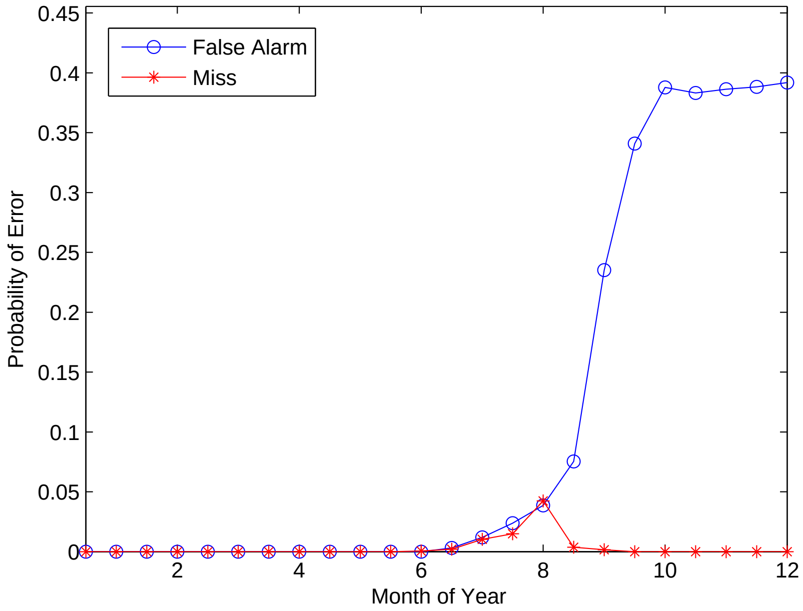

Figure A.13: *Errors prediction for 2005 based on observations from 2006, 2008, and 2009.*

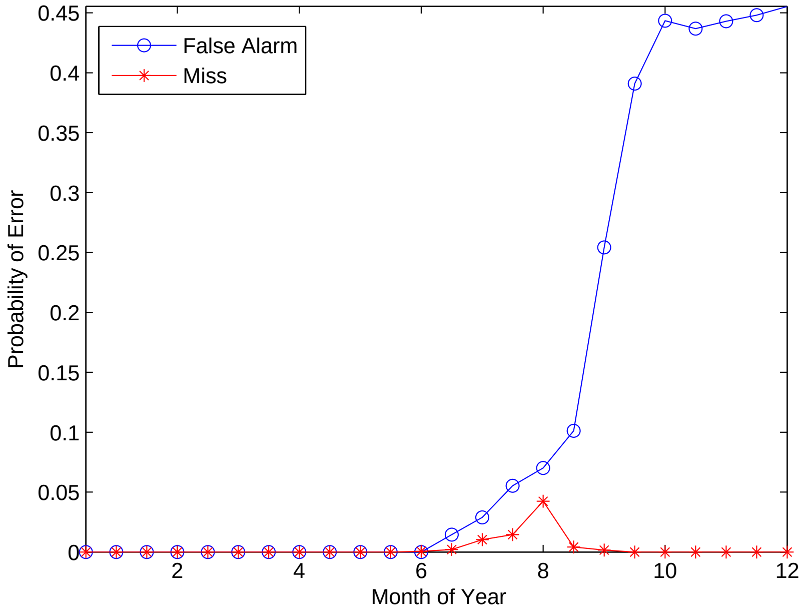

Figure A.14: *Errors prediction for 2005 based on observations from 2007, 2008, and 2009.*

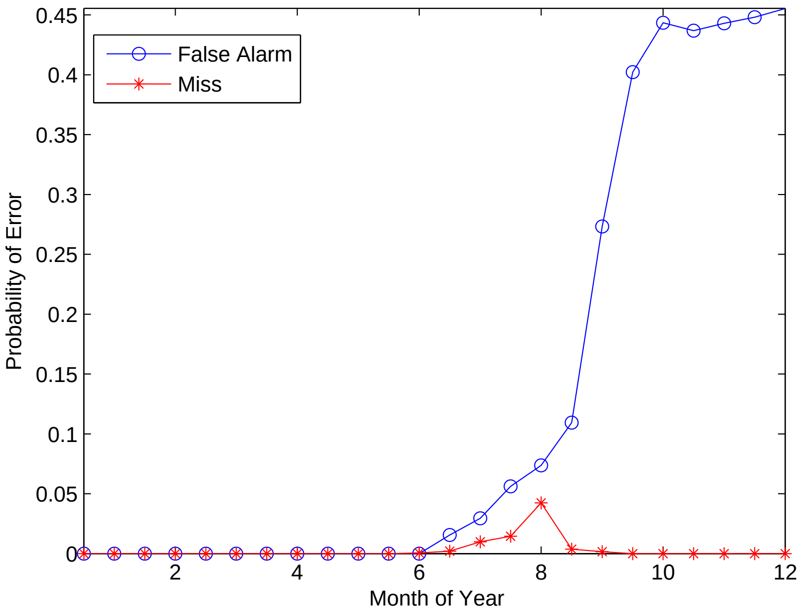

Figure A.15: *Errors prediction for 2005 based on observations from 4 other years.*

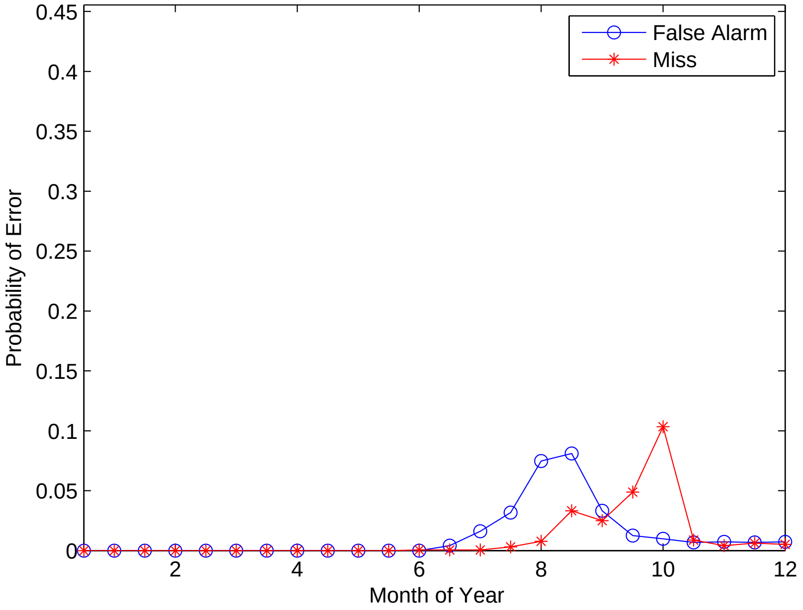

Figure A.16: *Errors prediction for 2006 based on observations from 2005.*

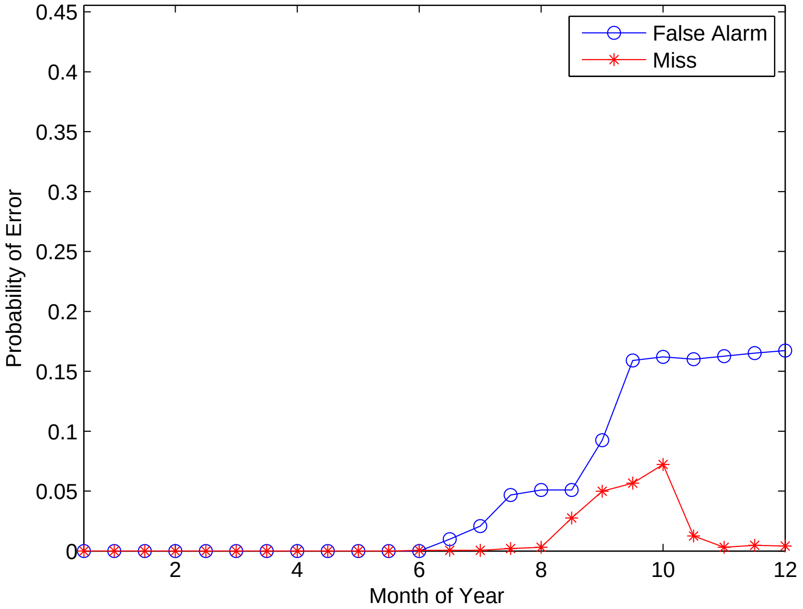

Figure A.17: *Errors prediction for 2006 based on observations from 2007.*

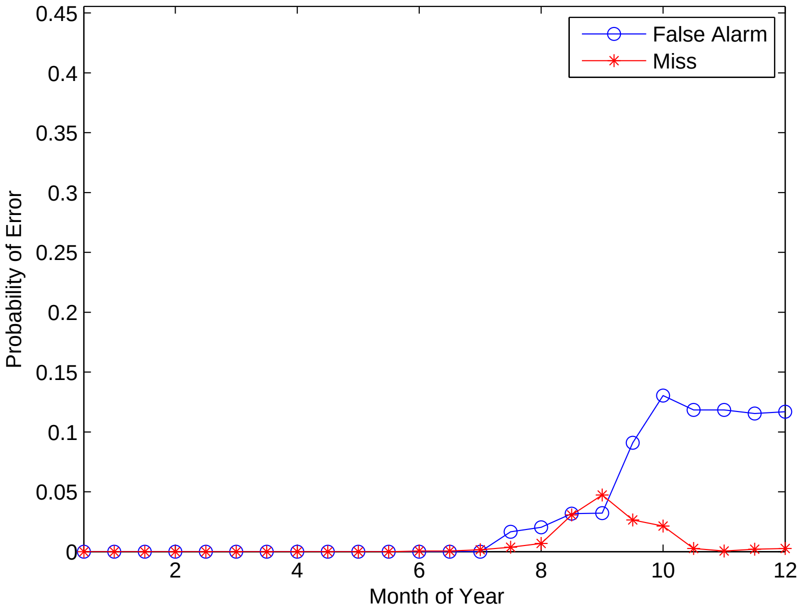

Figure A.18: *Errors prediction for 2006 based on observations from 2008.*

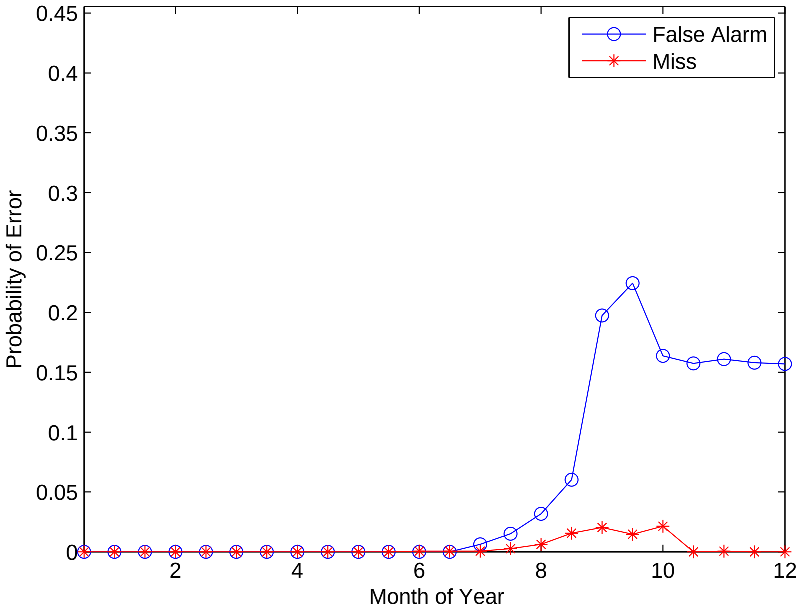

Figure A.19: *Errors prediction for 2006 based on observations from 2009.*

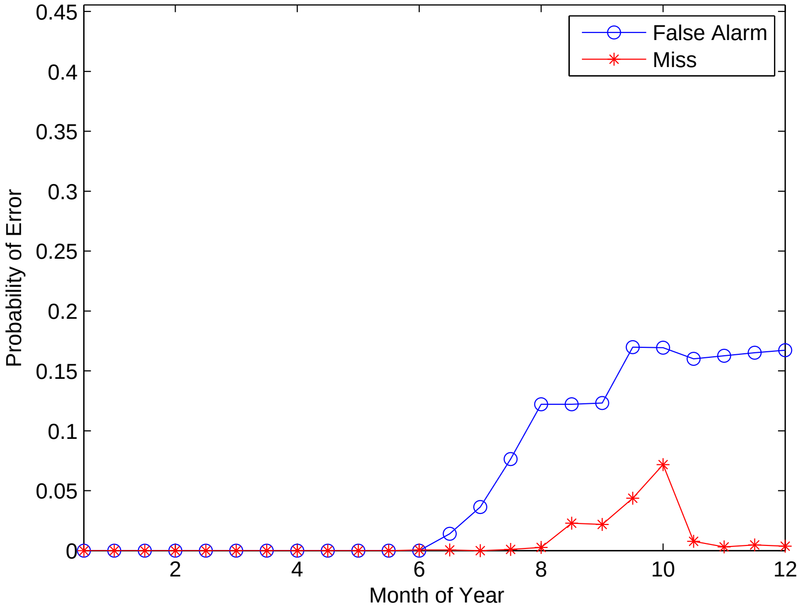

Figure A.20: *Errors prediction for 2006 based on observations from 2005 and 2007.*

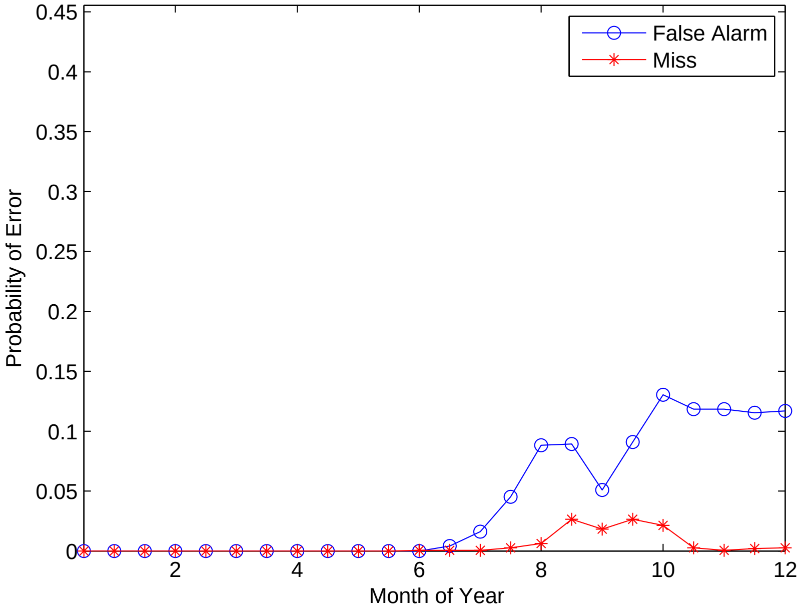

Figure A.21: *Errors prediction for 2006 based on observations from 2005 and 2008.*

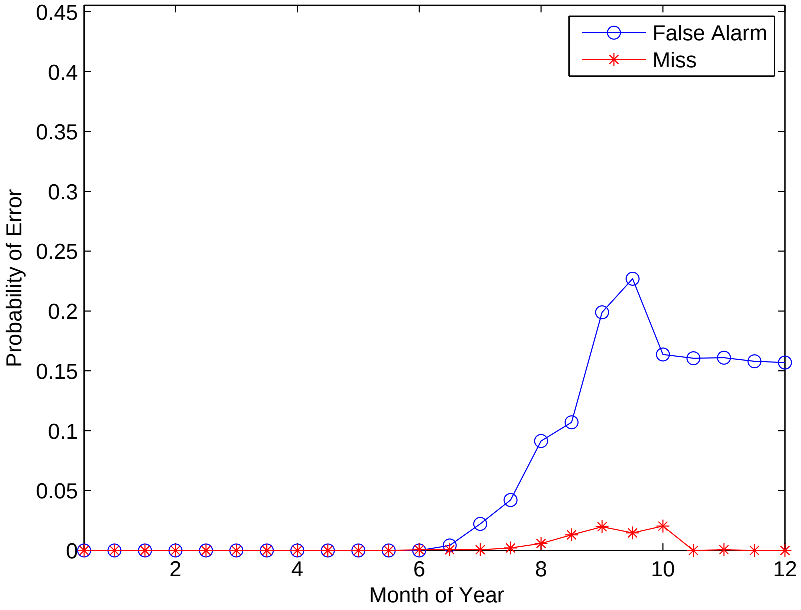

Figure A.22: *Errors prediction for 2006 based on observations from 2005 and 2009.*

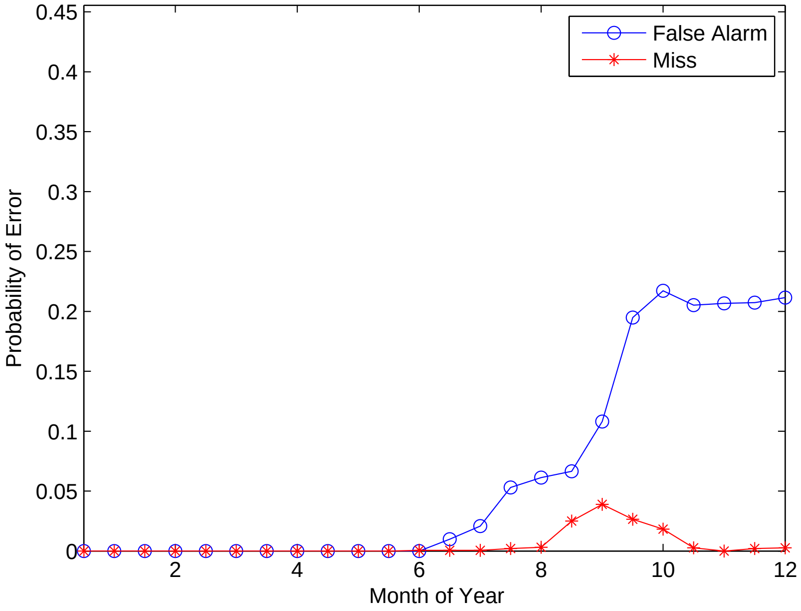

Figure A.23: *Errors prediction for 2006 based on observations from 2007 and 2008.*

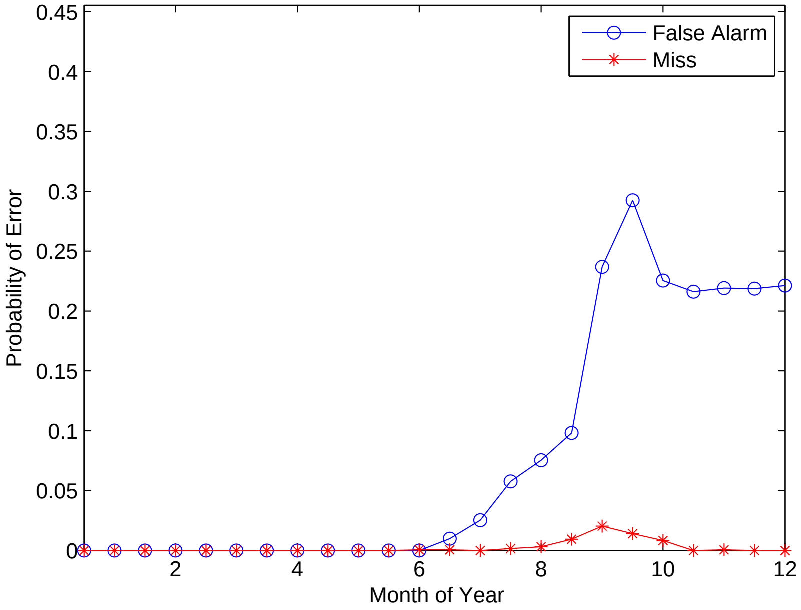

Figure A.24: *Errors prediction for 2006 based on observations from 2007 and 2009.*

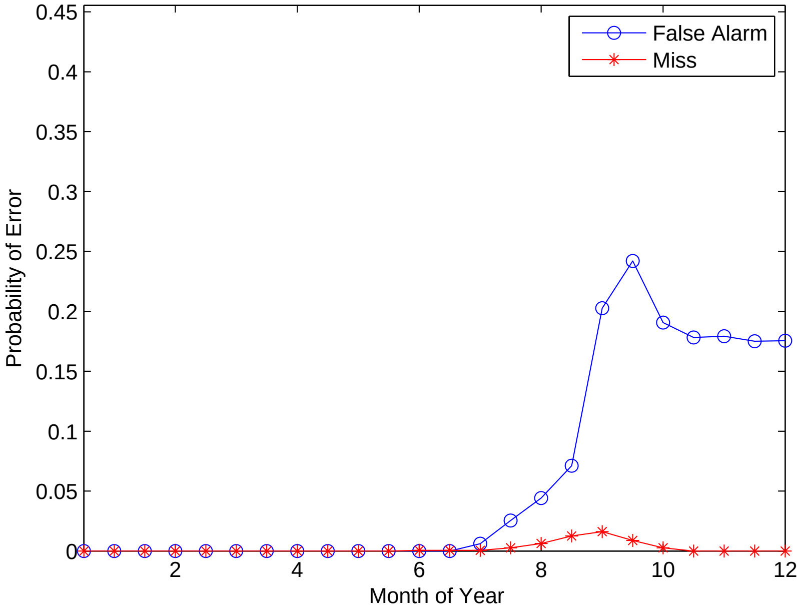

Figure A.25: *Errors prediction for 2006 based on observations from 2008 and 2009.*

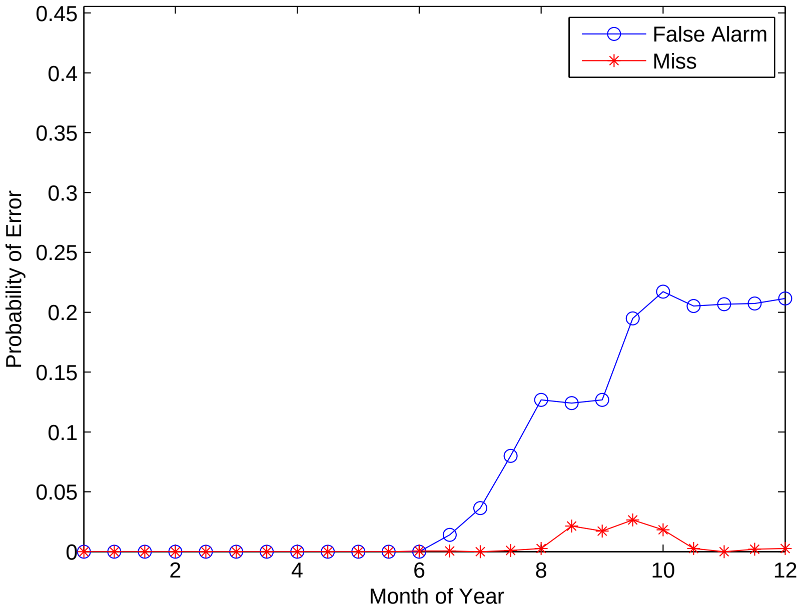

Figure A.26: *Errors prediction for 2006 based on observations from 2005, 2007, and 2008.*

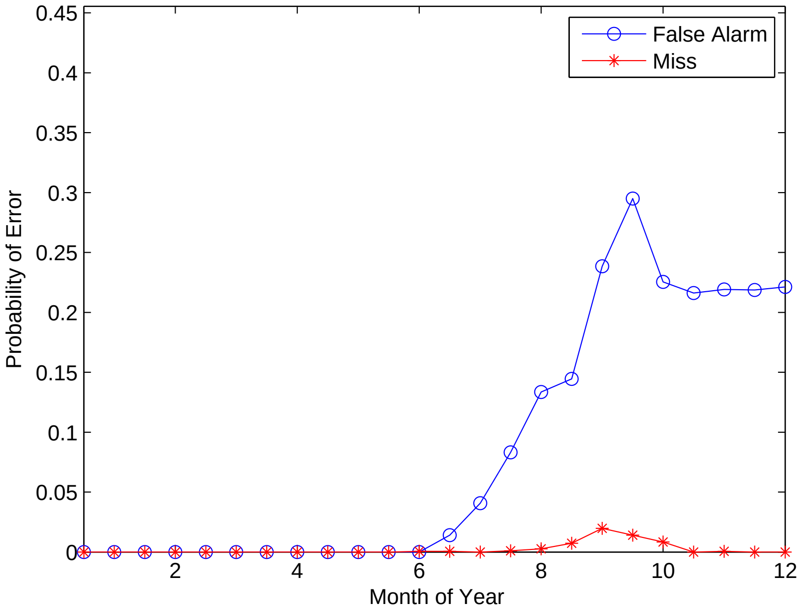

Figure A.27: *Errors prediction for 2006 based on observations from 2005, 2007, and 2009.*

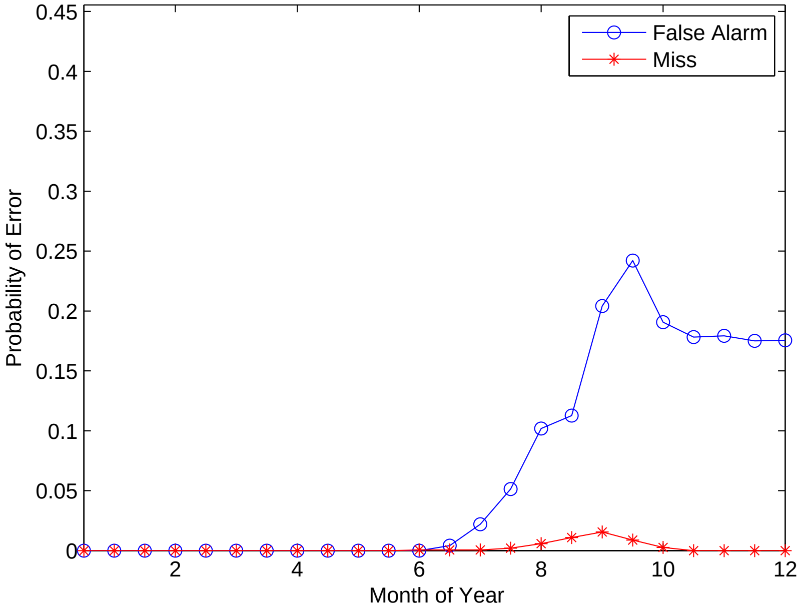

Figure A.28: *Errors prediction for 2006 based on observations from 2005, 2008, and 2009.*

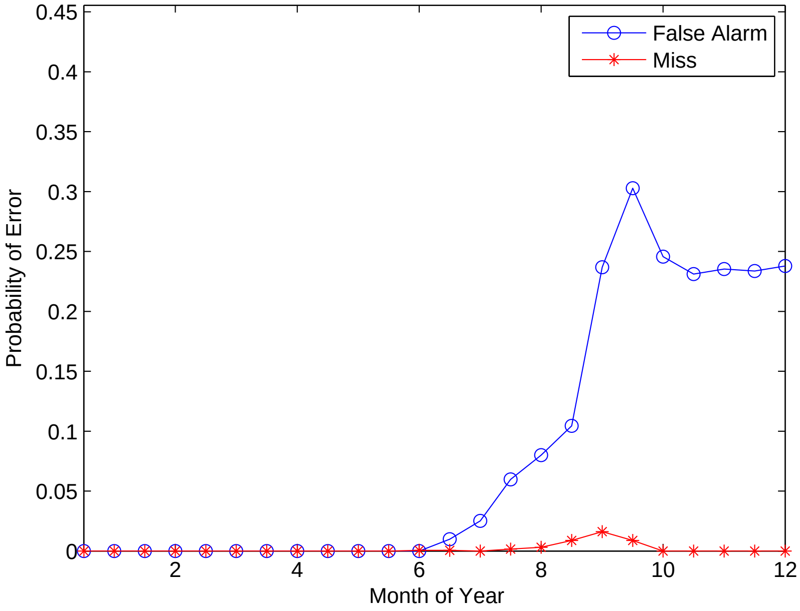

Figure A.29: *Errors prediction for 2006 based on observations from 2007, 2008, and 2009.*

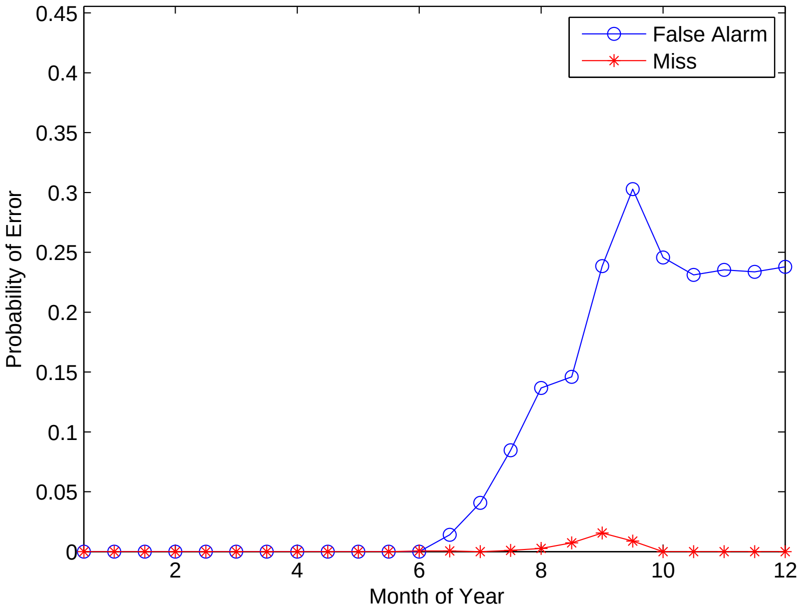

Figure A.30: *Errors prediction for 2006 based on observations from 4 other years.*

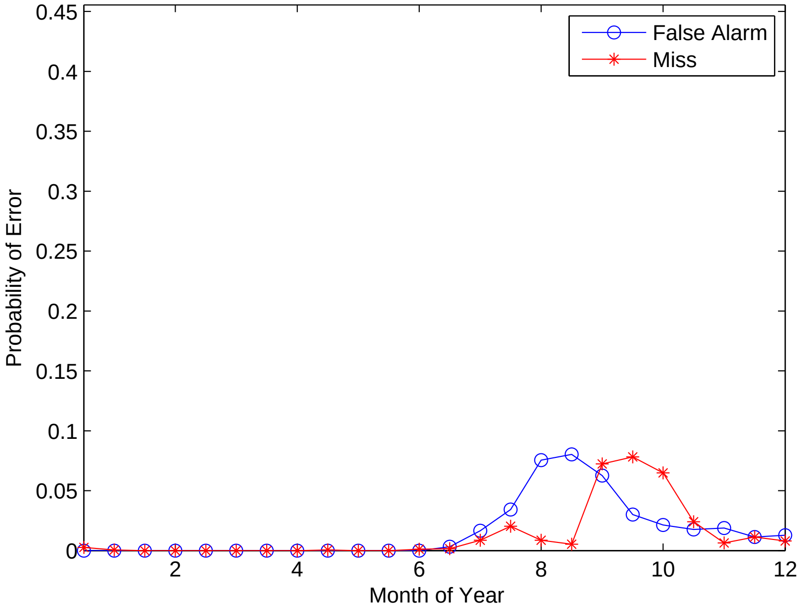

Figure A.31: *Errors prediction for 2007 based on observations from 2005.*

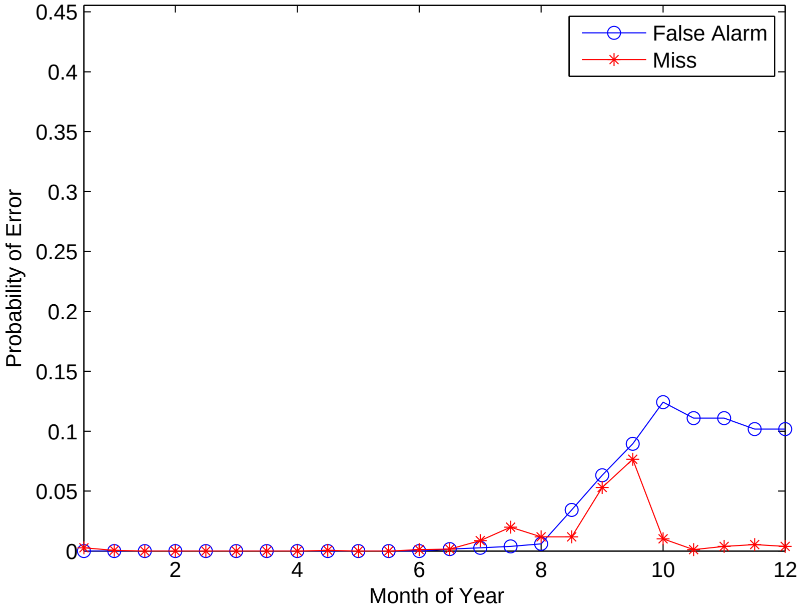

Figure A.32: *Errors prediction for 2007 based on observations from 2006.*

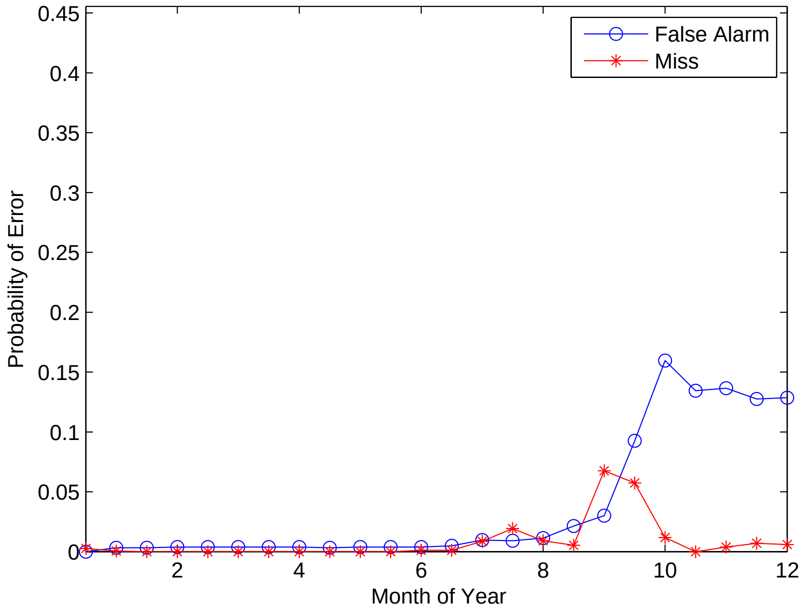

Figure A.33: *Errors prediction for 2007 based on observations from 2008.*

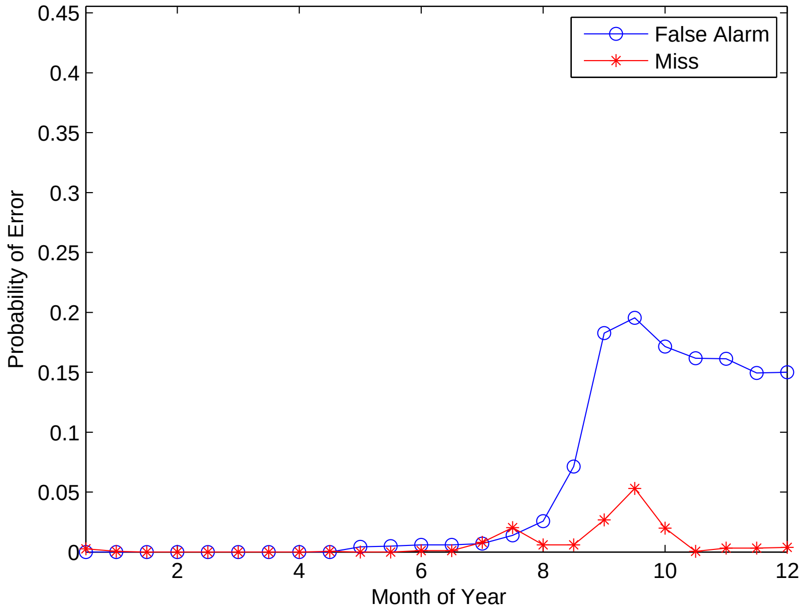

Figure A.34: *Errors prediction for 2007 based on observations from 2009.*

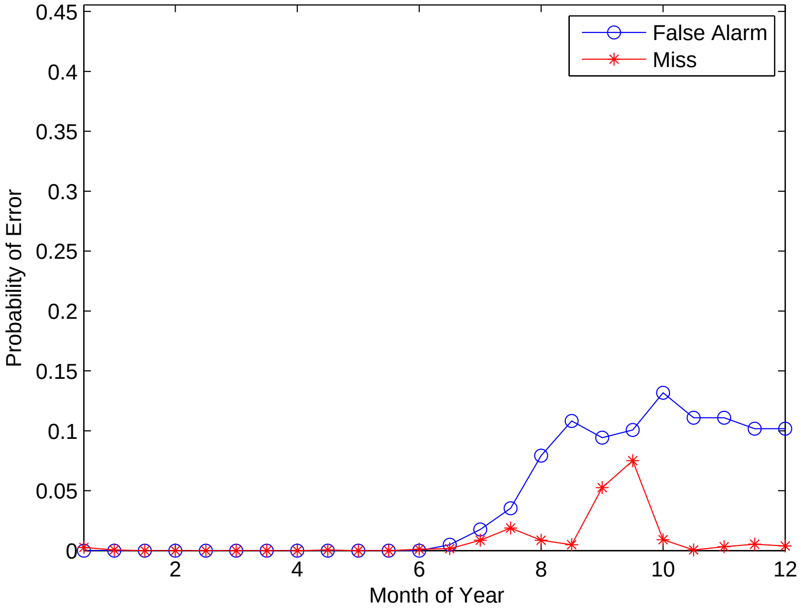

Figure A.35: *Errors prediction for 2007 based on observations from 2005 and 2006.*

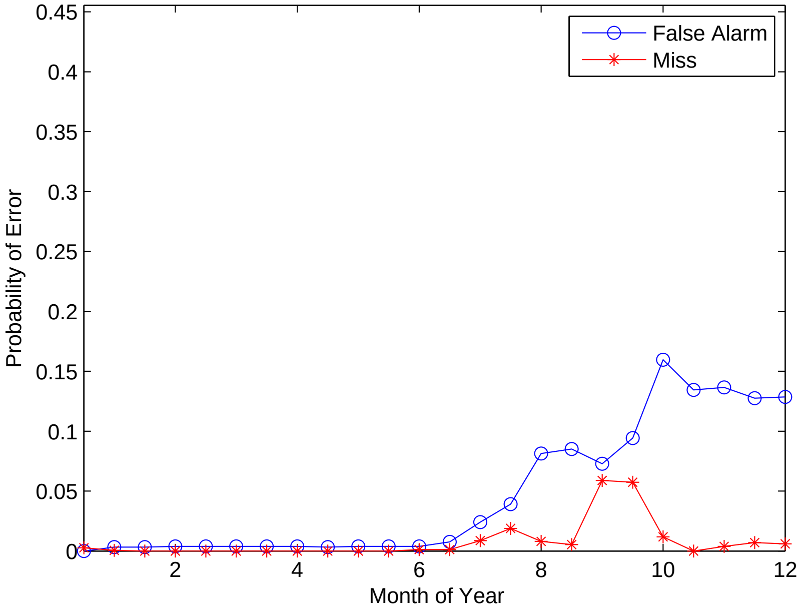

Figure A.36: *Errors prediction for 2007 based on observations from 2005 and 2008.*

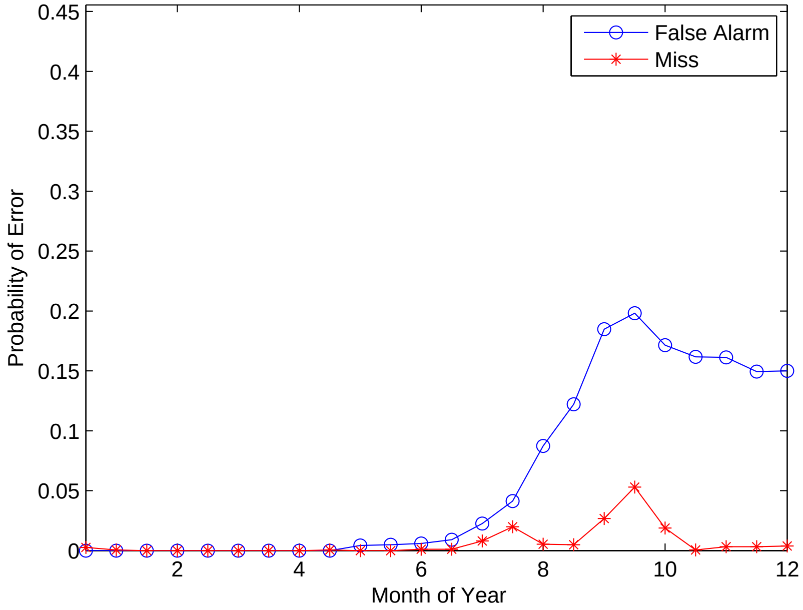

Figure A.37: *Errors prediction for 2007 based on observations from 2005 and 2009.*

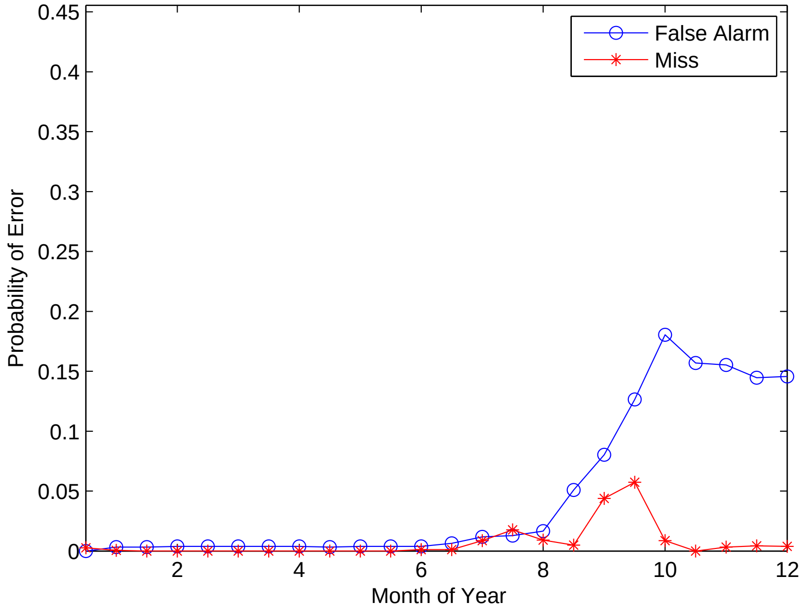

Figure A.38: *Errors prediction for 2007 based on observations from 2006 and 2008.*

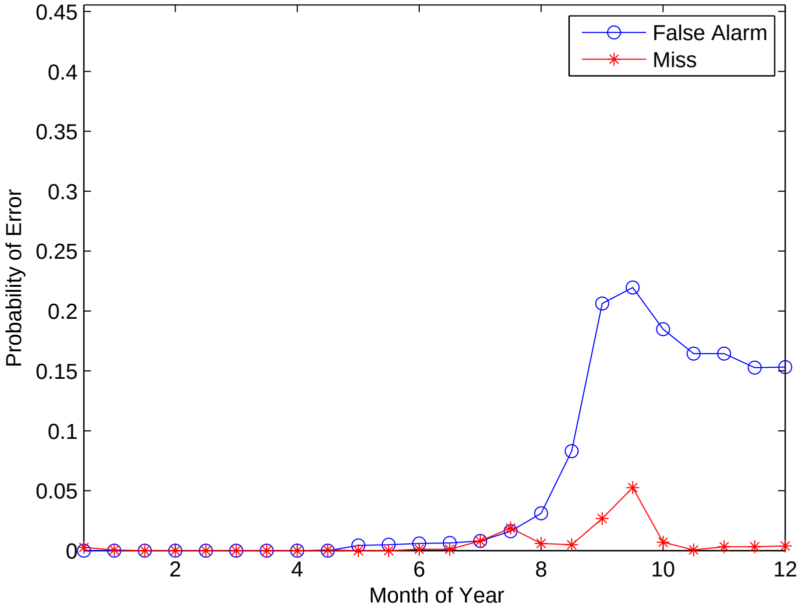

Figure A.39: *Errors prediction for 2007 based on observations from 2006 and 2009.*

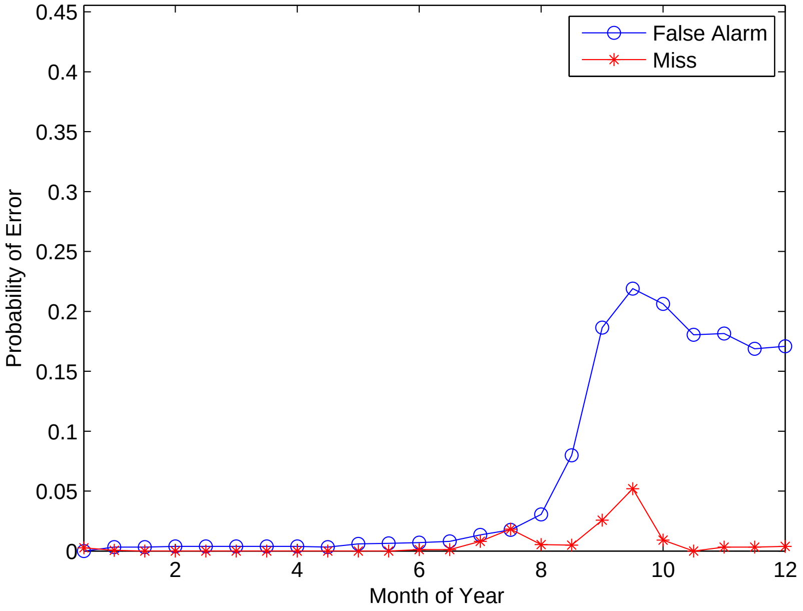

Figure A.40: *Errors prediction for 2007 based on observations from 2008 and 2009.*

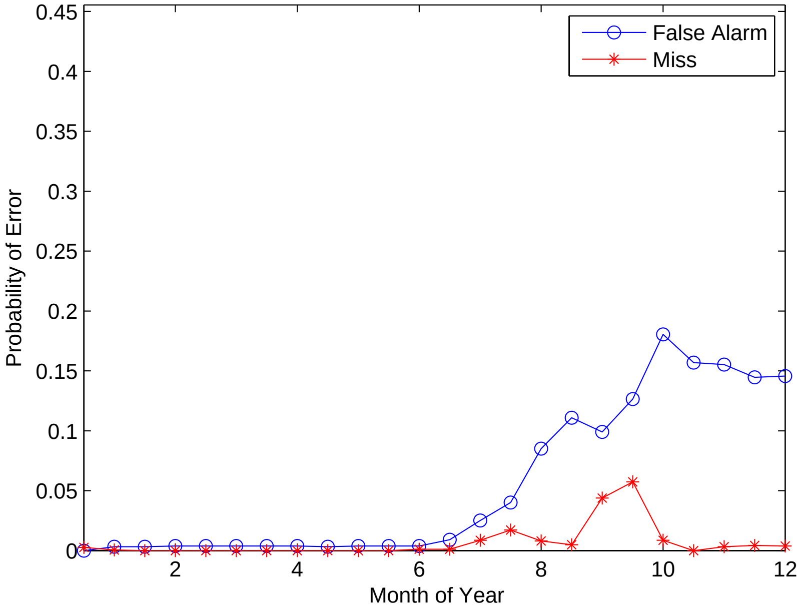

Figure A.41: *Errors prediction for 2007 based on observations from 2005, 2006, and 2008.*

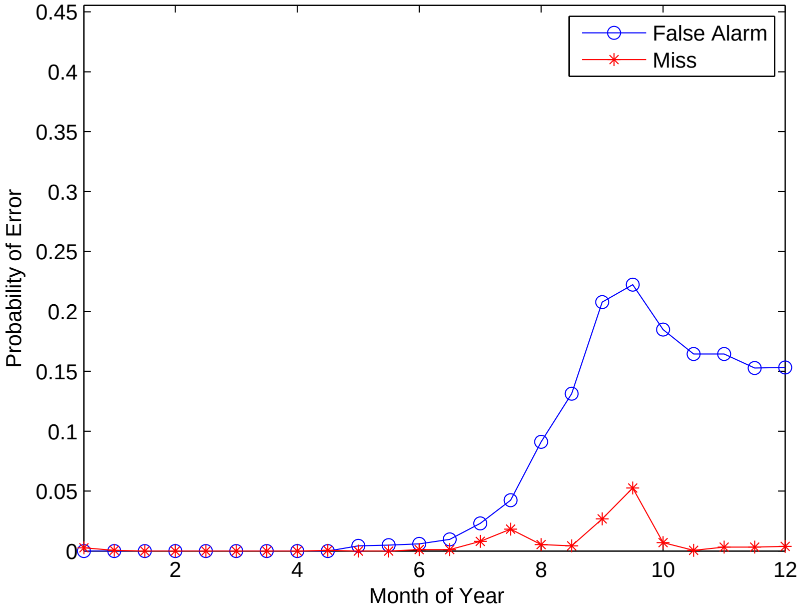

Figure A.42: *Errors prediction for 2007 based on observations from 2005, 2006, and 2009.*

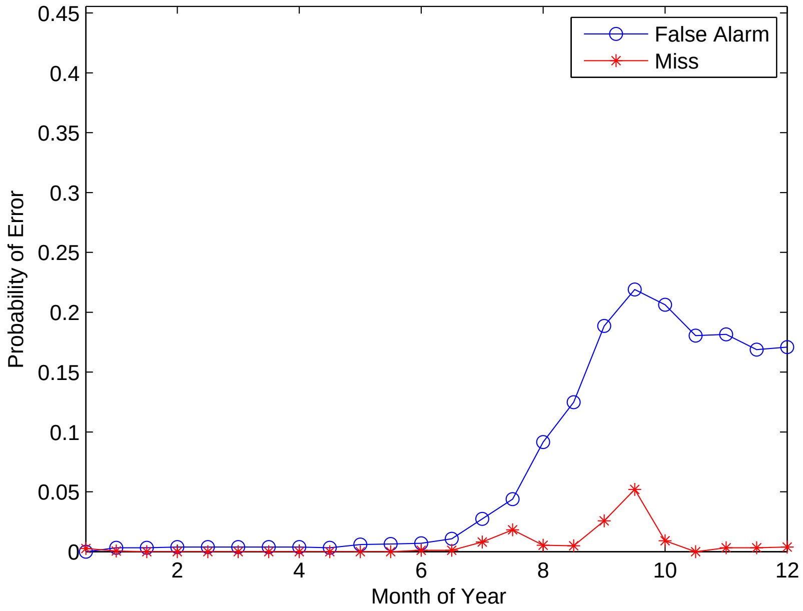

Figure A.43: *Errors prediction for 2007 based on observations from 2005, 2008, and 2009.*

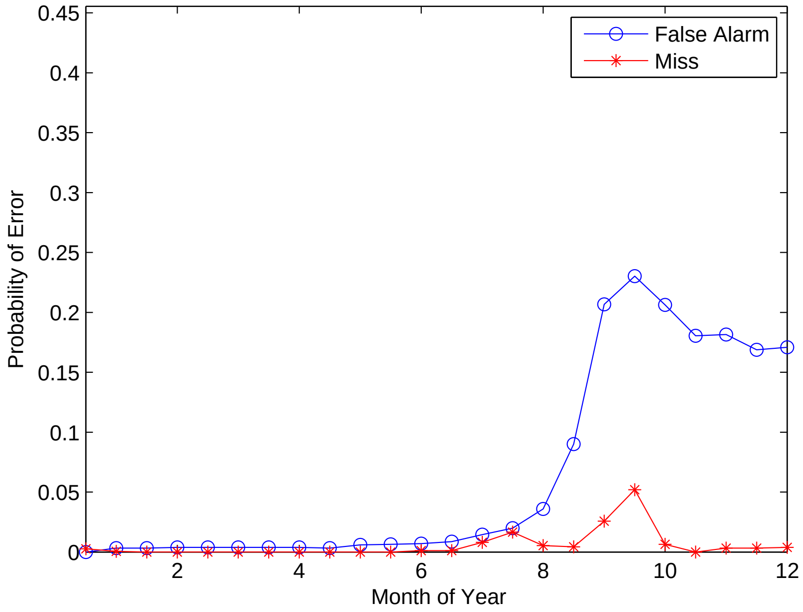

Figure A.44: *Errors prediction for 2007 based on observations from 2006, 2008, and 2009.*

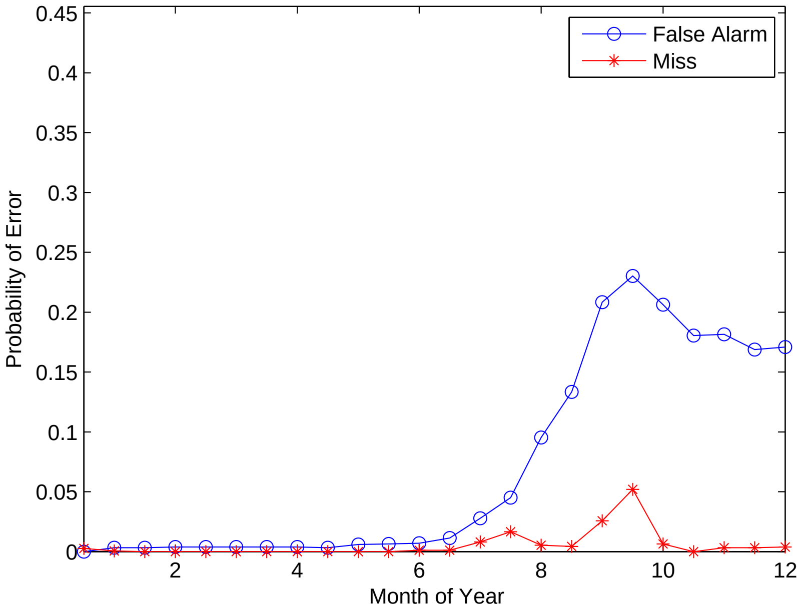

Figure A.45: *Errors prediction for 2007 based on observations from 4 other years.*

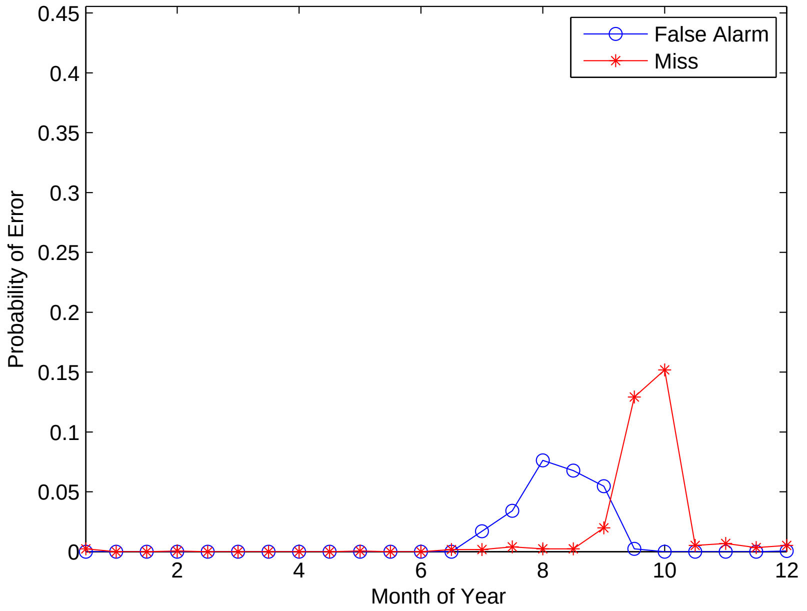

Figure A.46: *Errors prediction for 2008 based on observations from 2005.*

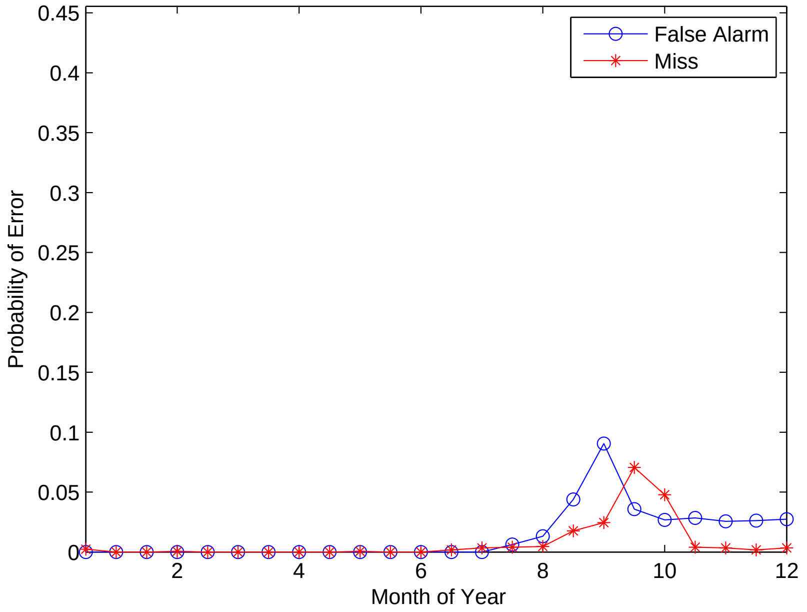

Figure A.47: *Errors prediction for 2008 based on observations from 2006.*

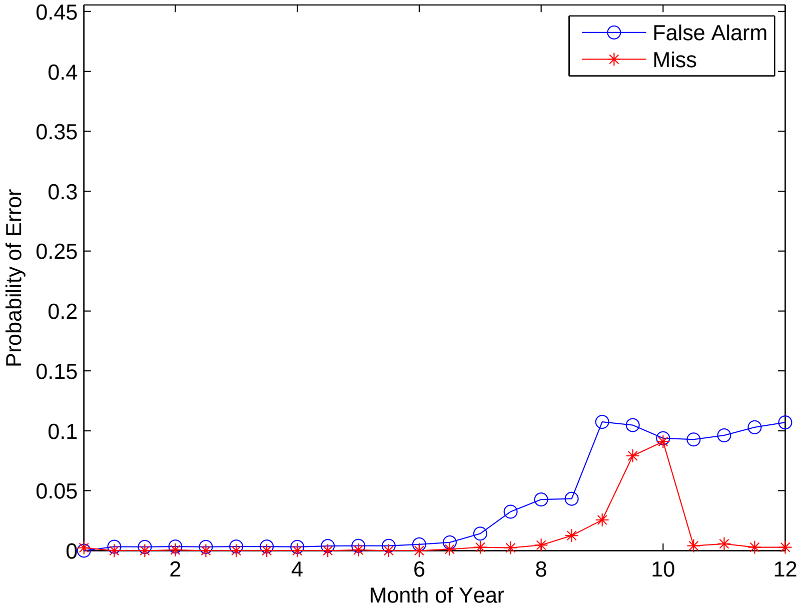

Figure A.48: *Errors prediction for 2008 based on observations from 2007.*

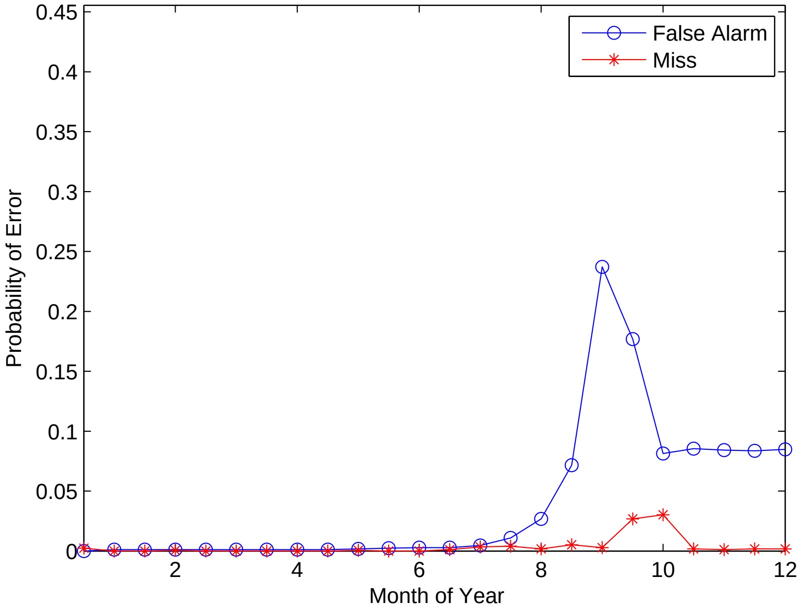

Figure A.49: *Errors prediction for 2008 based on observations from 2009.*

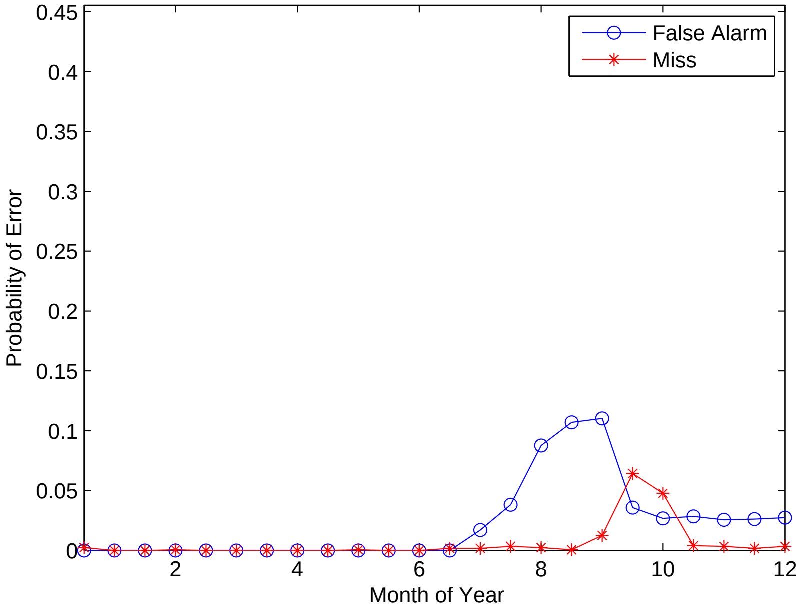

Figure A.50: *Errors prediction for 2008 based on observations from 2005 and 2006.*

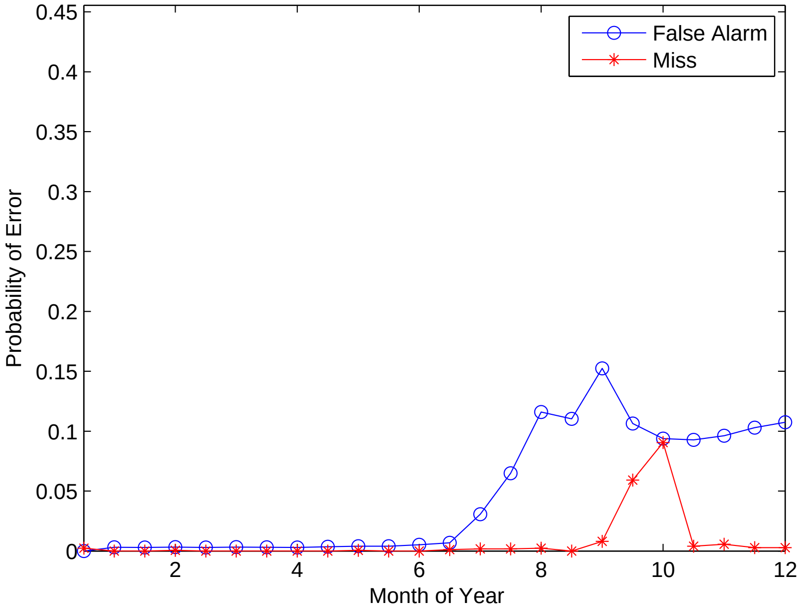

Figure A.51: *Errors prediction for 2008 based on observations from 2005 and 2007.*

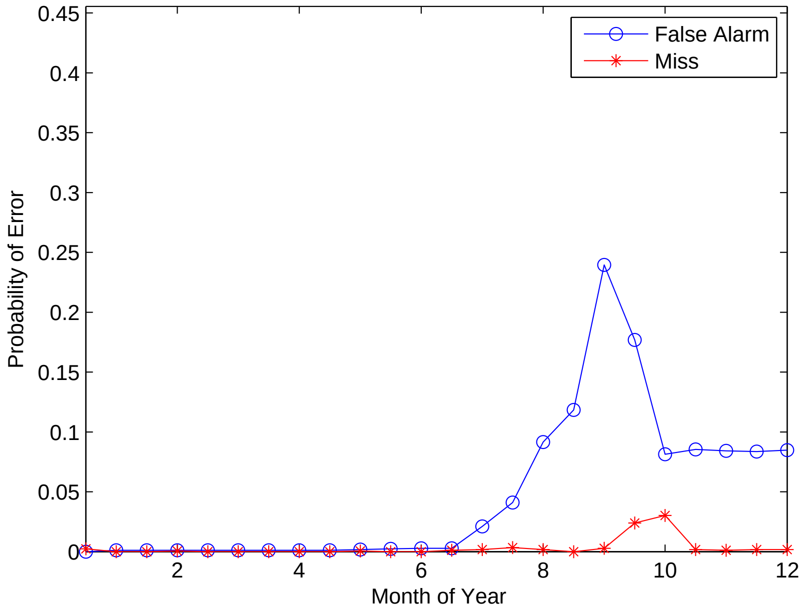

Figure A.52: *Errors prediction for 2008 based on observations from 2005 and 2009.*

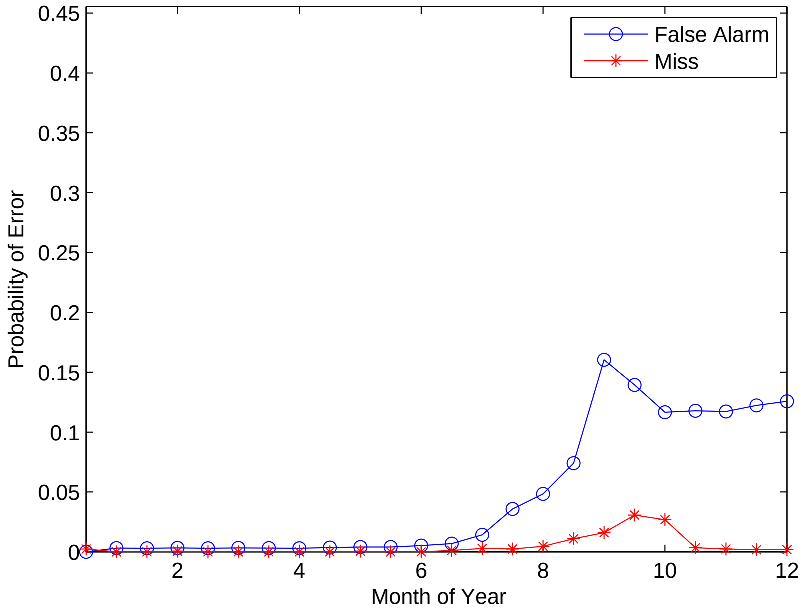

Figure A.53: *Errors prediction for 2008 based on observations from 2006 and 2007.*

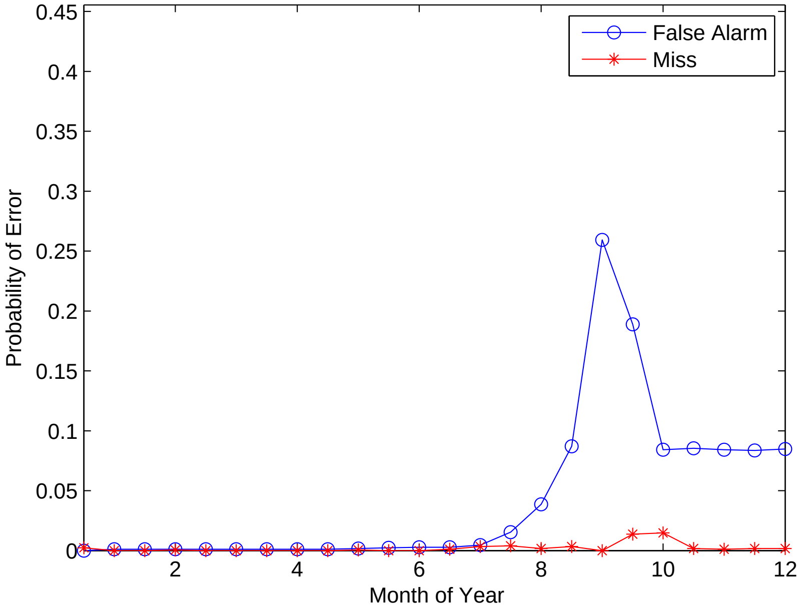

Figure A.54: *Errors prediction for 2008 based on observations from 2006 and 2009.*

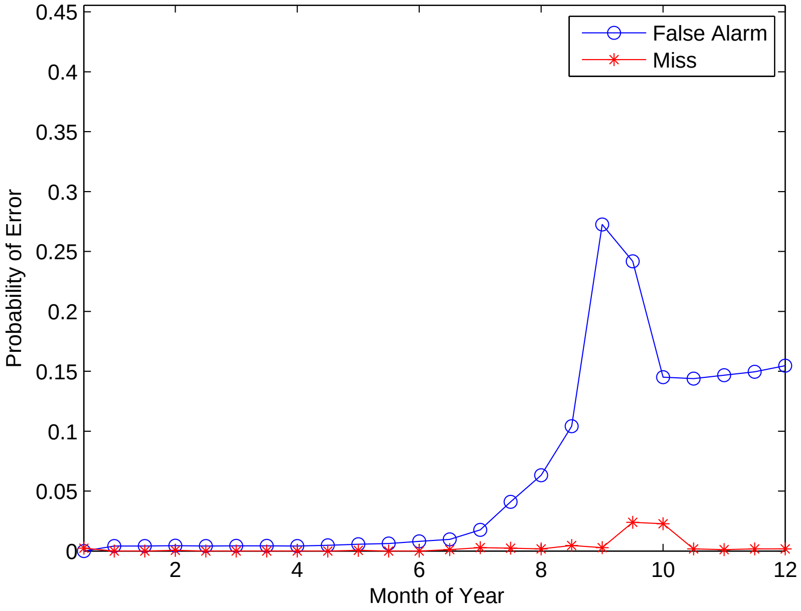

Figure A.55: *Errors prediction for 2008 based on observations from 2007 and 2009.*

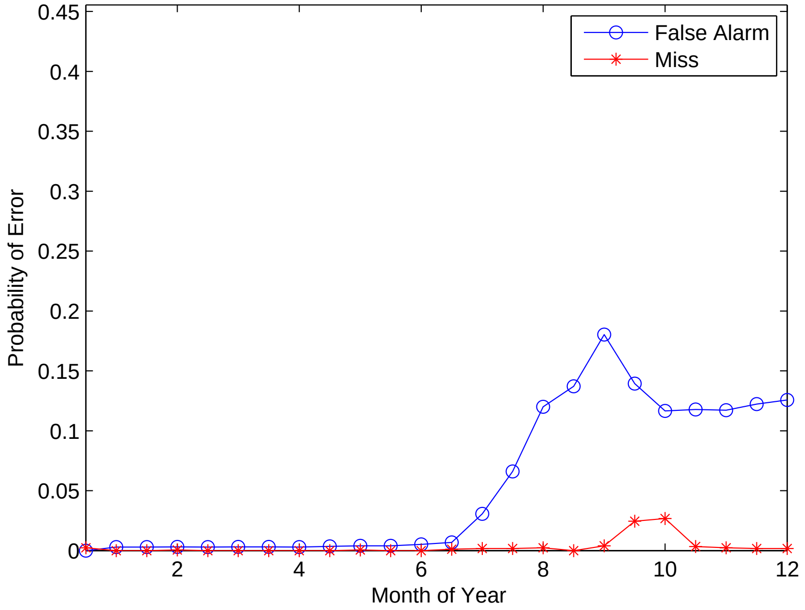

Figure A.56: *Errors prediction for 2008 based on observations from 2005, 2006, and 2007.*

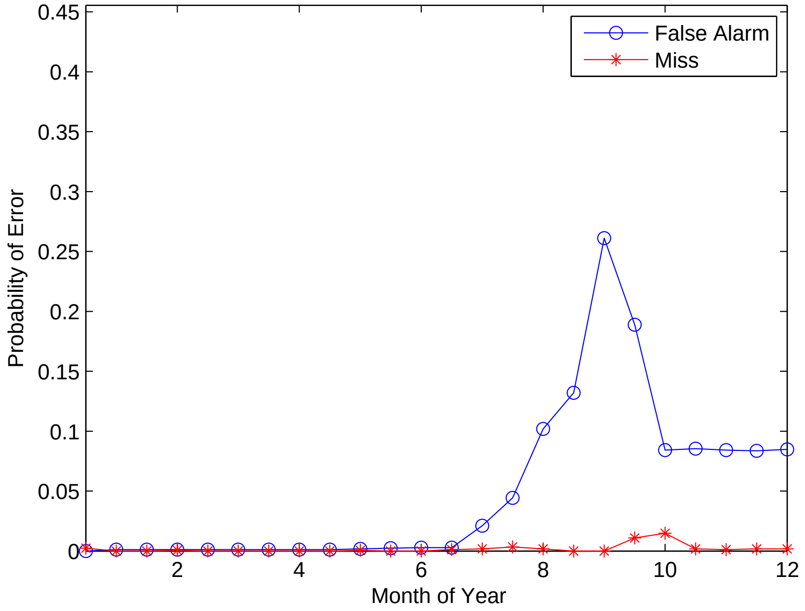

Figure A.57: *Errors prediction for 2008 based on observations from 2005, 2006, and 2009.*

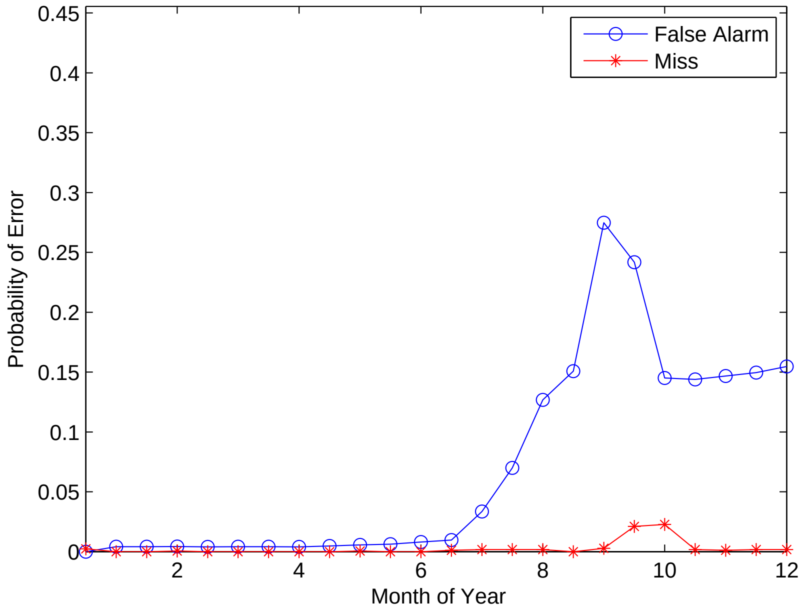

Figure A.58: *Errors prediction for 2008 based on observations from 2005, 2007, and 2009.*

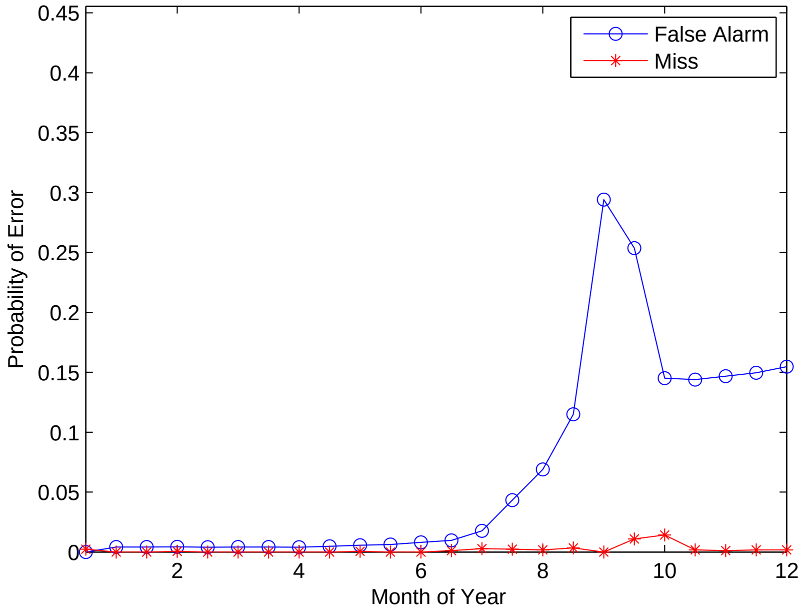

Figure A.59: *Errors prediction for 2008 based on observations from 2006, 2007, and 2009.*

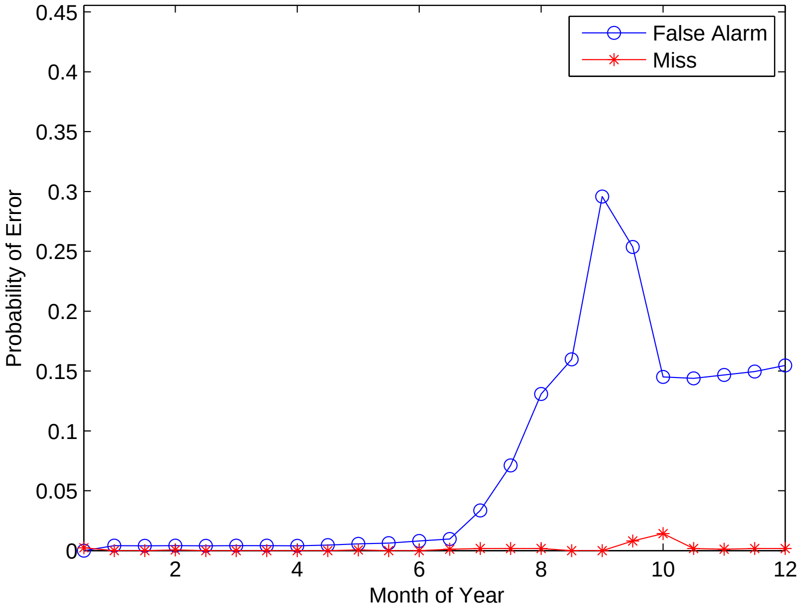

Figure A.60: *Errors prediction for 2008 based on observations from 4 other years.*